2345-324550967-73211208932

# Guía de Auditoría Algorítmica

Enero de 2021

# ÍNDICE DE CONTENIDOS

# Prefacio

La presente Guía de Auditoría Algorítmica ha sido elaborada y adaptada por un equipo de investigación de Eticas Research and Consulting SL, bajo el encargo y la supervisión de la Agencia Española de Protección de Datos. La metodología aquí propuesta se ha desarrollado sobre la base de textos especializados en este campo y a la experiencia del equipo auditor de Eticas Research and Consulting, con la colaboración del Dr. Carlos Castillo de la Universitat Pompeu Fabra.

# RESUMEN

Esta Guía de Auditoría Algorítmica ofrece directrices y orientaciones metodológicas para la realización de auditorías de aquellos productos y servicios enmarcados en el ámbito de la Inteligencia Artificial (IA) que

incluyan el uso de algoritmos y que, en alguna etapa del proceso, recopilen o traten datos de carácter personal.

Los servicios de IA basados en el uso de algoritmos se extienden con rapidez tanto en el sector público como en el privado. No obstante, los algoritmos son a menudo definidos como "cajas negras" de código informático y datos, cuyos resultados se vuelven cada vez más impredecibles e incontrolables. Esto genera muchas preocupaciones acerca de su impacto ético, social, jurídico y, también, empresarial. El respeto por los derechos fundamentales a la privacidad y a la protección de los datos personales forma parte de estas preocupaciones. La etiqueta "algoritmo" engloba diversos tipos de sistemas, dependiendo de los datos que el mismo maneja, del tipo de funcionamiento interno que se establece para dicho sistema, o de los objetivos de su funcionamiento, entre otros.

Esta guía no pretende realizar una definición técnica exhaustiva de este tipo de tecnologías, ni establecer una metodología de auditoría específica para cada una de ellas. Su objetivo es exponer una metodología general que sirva como hoja de ruta para la auditoría de diversas aplicaciones algorítmicas. Por lo tanto, la Guía está especialmente dirigida a las y los responsables del uso de algoritmos, de tratamiento de datos y de la realización de estas auditorías, aunque también pretende ampliar el conocimiento del público general, cada vez más interesado en comprender estas cuestiones.

## PALABRAS CLAVE

Auditoría Algorítmica, Algoritmos, Inteligencia Artificial, Aprendizaje de Máquinas, Machine Learning, Datos Masivos, Big Data, RGPD, Protección de Datos Personales, Atribución de Responsabilidad, Rendición de Cuentas, Cumplimiento legal, Ética.

# I. INTRODUCCIÓN

El reciente y rápido desarrollo de las nuevas tecnologías de procesamiento de datos masivos (big data) y, en particular, de aquellas que se sirven de algoritmos y técnicas de Inteligencia Artificial (IA), tienen importantes implicaciones a nivel social, económico, jurídico y ético. El

auge de estas nuevas tecnologías, no obstante, se está produciendo en un marco pre-normativo, que no contribuye a que su desarrollo e implementación sean todo lo explicables, equitativas y éticas que sería deseable. Si entendemos que la eficiencia de una nueva tecnología depende también de cuánto y cómo le sirva al conjunto de las personas y al desarrollo social, estas carencias también influyen en una disminución de esa misma eficiencia tecnológica. En este escenario, la necesidad de regular el uso de soluciones y algoritmos de IA, es clara. Actualmente existen iniciativas y propuestas a nivel europeo y organismos y estructuras administrativas en España que contribuyen a establecer líneas de orientación al respecto, pero es necesario reforzar dicho marco normativo.

El uso de algoritmos aumenta de forma constante, tanto por parte tanto del sector público, como del privado, incluyendo ámbitos como el político, el legislativo, el tecnológico, el financiero, las telecomunicaciones, el sector salud, la fabricación, el transporte, la energía o la educación, por poner algunos ejemplos. No obstante, los algoritmos, especialmente aquellos de aprendizaje automático se convierten, a menudo, en un conjunto opaco de código informático y datos, lo cual dificulta que otras personas o entidades puedan comprender, predecir o controlar qué ocurre en su interior y cuáles serán las implicaciones de las operaciones llevadas a cabo. Por este motivo, se ha extendido la definición de los algoritmos como “cajas negras”. Esto implica que el uso de algoritmos puede afectar de manera indeseable a las personas, a grupos de personas o al conjunto de la sociedad, dando lugar a potenciales riesgos, a menudo relacionados con posibles sesgos del sistema y formas de discriminación, capaces de afectar a individuos o grupos sociales vulnerables. Dichas formas de impacto social se irán definiendo a lo largo de la Guía. Además, esta opacidad de los sistemas algorítmicos pone en tela de juicio el respeto a la privacidad y a la protección de datos personales. Como se verá a lo largo de esta guía, esto implica analizar los algoritmos, también, en el contexto social, económico

y cultural del que forman parte, y de acuerdo con la perspectiva de las personas a las que este afecta, directa o indirectamente.

En este escenario, las auditorías algorítmicas se presentan como una forma necesaria de hacer que esta tecnología sea más explicable, más transparente, más predecible y más controlable por la ciudadanía, las instituciones públicas y también las empresas, ya sea antes del desarrollo del sistema, durante su desarrollo o a posteriori. Contribuyen también a mejorar los mecanismos de atribución de responsabilidad y de rendición de cuentas de los sistemas algorítmicos. La metodología para auditar algoritmos, sin embargo, no es sencilla ni está completamente definida todavía, lo cual supone un desafío.

En este contexto complejo, Eticas Research and Consulting presenta esta Guía de Auditoría Algorítmica, con tres objetivos principales:

- El primero, y más general, es clarificar el vínculo entre la realización de auditorías algorítmicas y la salvaguarda de los derechos fundamentales a la privacidad y a la protección de datos personales, que recoge la Carta de los Derechos Fundamentales de la Unión Europea.

- El segundo es aportar claridad respecto al necesario marco normativo para los sistemas algorítmicos, contribuyendo a la correcta interpretación e implementación del Reglamento Europeo de Protección de Datos del Parlamento Europeo y del Consejo, y de su ampliación en los casos necesarios.

- El tercero, que representa el principal elemento de interés de esta Guía, es ofrecer directrices y orientaciones metodológicas para la realización de auditorías algorítmicas que permitan examinar estas tecnologías, de manera que sean diseñadas, desarrolladas y utilizadas de una forma aceptable desde el punto de vista jurídico, pero también

previsible, proporcional, deseable, sostenible y socialmente justa y responsable.

La presente Guía de Auditoría Algorítmica, por tanto, se encuadra en lo propuesto por la Guía de Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial de la AEPD, en lo que respecta al cumplimiento efectivo de los principios de protección de datos personales y cómo el correcto planteamiento y desarrollo de una auditoría algorítmica puede contribuir a este objetivo. Esto es importante, dado que una auditoría algorítmica, puede tener, a su vez, un impacto indeseable desde el punto de vista social, jurídico, político o empresarial, si se realiza de manera inadecuada, dado que esta puede suponer una reconfiguración o un cambio en la implementación de un algoritmo, que resulte más perjudicial que la anterior. La auditoría algorítmica requiere, además, una especial atención en la recopilación y el tratamiento de los datos personales y sensibles involucrados en el análisis del algoritmo.

---

Para más información, se sugiere consultar las siguientes fuentes: *Artificial Intelligence for Europe*, *A European strategy for data*, *Commission Report on safety and liability implications of AI, the Internet of Things and Robotics*, *White Paper on Artificial Intelligence: a European approach to excellence and trust*, *Ethics Guidelines for Trustworthy AI*.

## 1.1 QUÉ ALGORITMOS DEBEN SER AUDITADOS

En esta guía, se utiliza la palabra algoritmo desde su concepción simple y actual en el ámbito de la ciencia de la computación, que es la más extendida. Desde esta perspectiva, un algoritmo consiste, básicamente, en un conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas que permite, típicamente, contestar una pregunta, tomar una decisión, solucionar un problema, realizar un cómputo, procesar datos o llevar a cabo alguna tarea.

Existen diversos tipos de algoritmos, tanto dependiendo de su modo de funcionamiento, como de los objetivos que persiguen. Dada esta dificultad, esta guía establece una metodología de auditoría general, replicable, con el afán de servir como hoja de ruta para que otros la apliquen a diferentes casos concretos. El foco principal de esta metodología está en detectar, prevenir y ayudar a corregir posibles consecuencias indeseables derivadas del uso de algoritmos.

La metodología de auditoría que se presenta en este manual está especialmente pensada para analizar aquellos algoritmos que puedan tener un impacto negativo sobre personas o grupos sociales, especialmente sobre aquellos en situaciones más vulnerables. Se considera que será especialmente importante auditar aquellos algoritmos que puedan afectar al acceso a la educación, al trabajo, a prestaciones o beneficios sociales, y/o que se implementen en ámbitos judiciales, de salud pública u otros ámbitos públicos y de relevancia e interés social.

Cualquier algoritmo debe desarrollarse e implementarse de manera que pueda ser auditado. No obstante, aquellos algoritmos considerados de impacto social, suponen un mayor riesgo para la protección de los

datos personales y para la privacidad e integridad de las personas. En la sección III de esta guía se propone una definición de aquellos tipos de impacto social, sesgo y discriminación en los que puede incurrir un algoritmo, y que deben evitarse. Por ejemplo, mientras un algoritmo utilizado para clasificar materiales en una línea de montaje es relevante desde el punto de vista operacional o económico, pero no es de interés desde una perspectiva de impacto social. En cambio, los algoritmos de selección de personal tienen diversas implicaciones para los derechos de los y las trabajadoras y, como ya se ha advertido, pueden derivar en violaciones de derechos como la discriminación por razones de género.

Por otra parte, para que un algoritmo de estas características pueda ser auditado con garantías de calidad, debe cumplir una serie de requisitos mínimos que se detallan en el apartado de Metodología de este manual. Esto es lo que consideraremos un "algoritmo auditable".

De acuerdo con la normativa vigente en materia de protección de datos (RGPD y LOPDGDD), todo tratamiento automatizado que produzca efectos significativos sobre la vida de una persona debe ser siempre supervisado por una persona. Esto implica el establecimiento claro de roles de responsabilidad relativos al desarrollo y el uso de un algoritmo, y también la obligación de establecer medidas de prevención y mitigación de riesgos. Para mejorar y reforzar el cumplimiento de estas medidas, la presente Guía recomienda que todo algoritmo utilizado en el sector público que cumpla los requisitos que se detallan en esta guía sea objeto de una auditoría algorítmica. Asimismo, los algoritmos utilizados en el sector privado, deberían tender a esta misma dinámica de manera progresiva y como parte de sus responsabilidades legales y sociales.

## 1.2 A QUIÉN SE DIRIGE ESTA GUÍA

Esta Guía de Auditoría Algorítmica se dirige, principalmente a aquellas personas en cuya tarea recae la responsabilidad del desarrollo y uso de algoritmos, y de su auditoría. Por lo tanto, se enfoca principalmente a las personas responsables de productos y de proyectos. No obstante, está pensada también para aportar un marco de comprensión estructurado a los equipos de carácter sociológico y técnico implicados en estos procesos, incluyendo aquellas personas: delegadas de protección de datos, responsables de ciberseguridad, cumplimiento ético y jurídico, personal técnico y equipos de desarrollo software y ciencia de datos. Por último, también pretende ampliar el conocimiento del público general, cada vez más interesado en comprender estas cuestiones.

# II. LA AUDITORÍA ALGORÍTMICA EN EL CONTEXTO DE LAS NORMAS RELATIVAS A LA PROTECCIÓN DE DATOS

Como se ha avanzado en el apartado de introducción, esta Guía de Auditoría Algorítmica se centra en desarrollar una metodología de auditoría, especialmente pensada para aquellos algoritmos cuyo desarrollo o implementación pueda tener un impacto social que afecte de manera particular a la protección de datos y a la privacidad de las personas.

Este apartado tiene dos objetivos principales. El primero es explicar cómo puede afectar el desarrollo y el uso de **algoritmos con impacto social a la protección de datos personales**. El segundo es explicar **cuál es la normativa vigente relacionada con la realización de auditorías algorítmicas**, e indicar cómo estas auditorías pueden ayudar al cumplimiento efectivo de dichas normas. En esta línea, se busca también hacer un mapeo de los textos vigentes, que ubique los conceptos utilizados en la metodología de auditoría algorítmica en el marco de la protección de datos.

Los algoritmos, especialmente aquellos que incorporan técnicas de aprendizaje artificial, pueden integrar y tratar cantidades masivas de datos, incluyendo datos de carácter personal y sensible. No obstante, como se ha señalado de forma recurrente en los últimos tiempos, los algoritmos a menudo tienen un diseño y un funcionamiento particularmente complejos y opacos, que impiden conocer y controlar cómo se tratan esos datos. Al mismo tiempo, se ha demostrado que el análisis extensivo de datos puede revelar información de carácter sensible, que los datos no mostraban de forma aislada. A esto se suma

que la finalidad y la utilidad de estos sistemas no siempre se comunica de forma clara y transparente, mientras que los algoritmos se utilizan de manera cada vez más frecuente, para sustituir tareas hasta el momento realizadas por humanos, que incluyen la ordenación, la predicción, la recomendación o el acompañamiento en la toma de decisiones, entre otras.

El desarrollo de nuevas técnicas de recopilación y tratamiento de datos masivos en las últimas décadas, ha dado lugar a un refuerzo en las normas éticas y jurídicas al respecto de la privacidad y la protección de datos personales. No obstante, estas siguen siendo insuficientes para dar cumplimiento a estos derechos y no se desarrollan al mismo ritmo que las soluciones tecnológicas. Los derechos a la privacidad y a la protección de

datos personales están recogidos en diferentes documentos nacionales y comunitarios de la Unión Europea, como derechos de carácter fundamental. En concreto, la Carta de los Derechos Fundamentales de la Unión Europea los refleja en sus Artículos 7 y 8 de la siguiente manera:

- **Artículo 7. Respeto de la vida privada y familiar:**

  *Toda persona tiene derecho al respeto de su vida privada y familiar, de su domicilio y de sus comunicaciones.*

- **Artículo 8. Protección de datos de carácter personal**

  *Toda persona tiene derecho a la protección de los datos de carácter personal que le conciernan. Estos datos se tratarán de modo leal, para fines concretos y sobre la base del consentimiento de la persona afectada o en virtud de otro fundamento legítimo previsto por la ley. Toda persona tiene derecho a acceder a los datos recogidos que le conciernan y a obtener su rectificación. El respeto de estas normas estará sujeto al control de una autoridad independiente.*

La especial importancia de estos derechos en lo que respecta al uso de algoritmos que tratan datos personales, y especialmente aquellos que lo hacen de manera extensiva, pone de manifiesto la necesidad de establecer medidas de control efectivo, de corrección, responsabilidad, rendición de cuentas y transparencia relativas al tratamiento de los datos. Por ello, esta Guía ofrece **directrices y orientaciones metodológicas para la realización de auditorías algorítmicas**, que permitan analizar e identificar los puntos de tensión que pueden suponer un incumplimiento de la normativa de protección de datos. Dichas auditorías permiten detectar posibles sesgos o malas prácticas en el procesamiento automático de datos, de cara a corregirlos y tenerlos en cuenta como requisitos de diseño en el desarrollo y uso de algoritmos y soluciones de IA. Esto implica desarrollar mecanismos que permitan examinar estas tecnologías, para contribuir a que sean diseñadas, desarrolladas y

utilizadas de una forma aceptable desde el punto de vista jurídico, pero también previsible, proporcional, deseable, sostenible y socialmente justa y responsable.

Por lo que respecta al cumplimiento de la normativa jurídica, que nos ocupa en este apartado, hay que señalar que, desde el 25 de mayo de 2018, es directamente aplicable en los estados miembros de la Unión Europea el **Reglamento 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en cuanto al tratamiento y la libre circulación de datos personales**[2] (en adelante, RGPD). La transposición española de este Reglamento europeo, ha dado lugar a la elaboración y entrada en vigor de la **Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales**[3] (en adelante, LOPDGDD).

[2] El Reglamento General de Protección de Datos se puede consultar en esta página web: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679.

[3] La Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales se puede consultar en esta página web: https://www.boe.es/eli/es/lo/2018/12/05/3.

Por su parte, la **Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales** (LOPDGDD) complementa y particulariza lo dispuesto por el Reglamento en el caso español, reforzando la importancia de dar cumplimiento a los principios de protección de datos y a la atención del ejercicio de derechos por parte del responsable, a la vez que incluye determinadas disposiciones aplicables a tratamientos concretos, algunos de los cuales pueden apoyarse en el desarrollo de soluciones que hagan uso de algoritmos.

Tanto el **RGPD** como la **LOPDGDD** vienen a establecer los principios directores que cualquier tipo de tratamiento, incluidos aquellos basados en soluciones de Inteligencia Artificial y que utilicen algoritmos, debe respetar definiendo un marco de desarrollo de las actuaciones de los responsables basado en la gestión de los riesgos para los derechos y las libertades de los interesados y la rendición de cuentas, o capacidad de demostrar el cumplimiento de las obligaciones impuestas por la normativa.

Este enfoque, **exige a responsables y encargados del tratamiento atender estos requerimientos de manera proactiva**, también en el caso de los tratamientos basados en un procesamiento automático de datos personales. Este marco regulatorio debe ser tenido en cuenta por las organizaciones para para evitar duplicación innecesaria de responsabilidades o promover obligaciones contradictorias, así como para no favorecer la ambigüedad e inseguridad jurídica en distintos ámbitos sectoriales.

# III. METODOLOGÍA DE LA AUDITORÍA ALGORÍTMICA

La auditoría se concentra en analizar e identificar aquellos aspectos del diseño, desarrollo e implementación de un algoritmo, que pueden suponer producir un impacto desventajoso para los grupos desaventajados y un incumplimiento de la normativa de protección de datos de cara a corregirlos y tenerlos en cuenta como requisitos de diseño en el desarrollo y uso de soluciones de IA. Con ello, busca contribuir a que estos algoritmos sean diseñados, desarrollados y utilizados de una forma **adecuada** desde el punto de vista jurídico, pero también que sean más **controlables,**

**deseables, sostenibles y socialmente justos y responsables**. Esto implica que tengan un tratamiento igualitario de los grupos sociales implicados, sean transparentes y accesibles por parte de la ciudadanía, e incorporen mecanismos de seguridad para prevenir, identificar y mitigar posibles sesgos. Establecer un marco general para el desarrollo de estas auditorías es fundamental, dado que una auditoría algorítmica implementada de forma inadecuada, puede tener también consecuencias indeseables, si no proponen medidas de corrección y mejora adecuadas, o no prestan especial atención a las propias medidas de recopilación y tratamiento de los datos personales y sensibles involucrados en el análisis del algoritmo.

Una auditoría algorítmica se compone de una serie de fases que convergen en un único objetivo: **identificar, anticipar y corregir** los posibles riesgos que surjan durante el ciclo de vida del algoritmo y de los datos tratados. Esto permite, a su vez, reforzar los mecanismos de **responsabilidad y rendición de cuentas** y de **protección de los derechos y libertades** de las personas físicas involucradas (ya sean personas individuales o grupos) y, especialmente los derechos fundamentales a la privacidad y a la protección de los datos personales.

Una auditoría algorítmica **puede ser interna o externa**. No obstante, una auditoría siempre debe contar con la colaboración de la persona o el equipo interno de la institución que implementa el algoritmo (o el cliente) y del equipo que lo desarrolla o ha desarrollado. La auditoría externa, si se realiza por una entidad confiable, con una experiencia y una preparación certificada, que aplique medidas adecuadas de seguridad de la información, y siga una metodología consolidada, puede resultar más objetiva.[4]

[4] Para no complicar innecesariamente el desarrollo metodológico de la auditoría algorítmica, no se hará referencia constante a esta distinción entre la auditoría interna y la externa a lo largo del documento.

## 3.1 OBJETIVOS GENERALES DE LA AUDITORÍA ALGORÍTMICA

La metodología propuesta, busca servir como una **garantía de calidad** de aquellos algoritmos de impacto social desarrollados e implementados por instituciones públicas y privadas, investigadores, emprendedores e innovadores. Asimismo, superar carencias en los procesos y las medidas de **responsabilidad** y de **rendición de cuentas**, relativas a las acciones derivadas del funcionamiento de los algoritmos. Esto implica establecer procedimientos de análisis de estos sistemas que conlleven, por una parte, un ejercicio de reflexión crítica y **concienciación** sobre su posible impacto y, por otra, la implementación de mecanismos de **transparencia** que permitan conocer los pasos del diseño y el desarrollo del sistema.

El fin de una auditoría es **identificar o anticipar errores, riesgos o amenazas** (actuales o potenciales) y **ayudar a corregirlos**. Esto puede darse en cualquiera de las fases del desarrollo del sistema, tanto en su diseño y puesta en marcha, como en la fase de funcionamiento y posteriormente a él. Por lo tanto, también permite trazar una estrategia de mejora de los procesos con intervención algorítmica en el futuro y responder a un fallo una vez que el algoritmo ha sido puesto en marcha. No obstante, cabe resaltar la importancia de implementar métodos de auditoría previos al despliegue y puesta en marcha de los sistemas. El sector tecnológico, incluidas las empresas e instituciones públicas, debe habituarse a auditar sus algoritmos, como una forma de asegurar su **responsabilidad social**. Como veremos, esto comparte muchos de los principios y resultados de las evaluaciones de protección de datos y de la privacidad.

Dependiendo de **quién realice esta auditoría**, los objetivos concretos de la auditoría podrán variar. Esto quiere decir que una auditoría

realizada con **objetivos de investigación**, permitirá generar conocimiento fundamental y aplicado sobre el comportamiento de los sistemas algorítmicos y sus efectos y reportarlo a la sociedad. En el caso de las auditorías desarrolladas por **organizaciones de la sociedad civil**, el objetivo podría ser investigar los sistemas que podrían afectar a las personas con las que trabajan o a las que defienden. En el caso de la **consultoría**, la auditoría podrá servir para recomendar mejoras en los sistemas desarrollados por instituciones públicas o privadas, para evitar que estos generen sesgos y formas de discriminación. Como último ejemplo, si estas son realizadas por la **misma institución que desarrolla o implementa el algoritmo**, servirán como una forma de autoevaluación de riesgos e impacto.

El tipo de evaluación que pueda llevar a cabo una auditoría, dependerá de la fase de desarrollo e implementación del algoritmo, o de su **ciclo de vida**. Esto quiere decir que, mientras en las primeras fases de la auditoría, se podrán realizar **análisis de los posibles riesgos**, en las últimas fases se podrán implementar medidas de **análisis de su impacto** real.

La metodología de auditoría algorítmica que aquí se propone, tiene en cuenta la importancia de que se realice una parte de **análisis técnico**, que permita evaluar la **eficacia** del sistema en sí mismo (y de acuerdo con los objetivos establecidos para este), y otra parte de **análisis cualitativo**.

Esta segunda parte de la auditoría tiene por objetivo valorar la **deseabilidad** y la **aceptabilidad** de un algoritmo, desde una perspectiva más amplia, que tenga en cuenta cómo este se **implementa**, se **integra** en su contexto social, a qué sistemas previos sustituye (si lo hace), qué nuevas dinámicas introduce, etc.

Cuando se audita un algoritmo, el objetivo es **ganar conocimiento sobre el sistema en sí mismo y sobre el entorno** (general y concreto) en el que se integra y opera este sistema. Esto implica preguntarse si su

funcionamiento es adecuado y pertinente, si cumple la legalidad vigente, si es eficaz, si es replicable en contextos similares y si es robusto; pero también implica preguntarse si es transparente, si es explicable, si es útil y si se utiliza de manera adecuada, o si es deseable desde un punto de vista ético, social y cultural. Esto debe permitir conocer si el modelo algorítmico puede haber sido diseñado sobre unas bases inestables o inapropiadas, o si su desarrollo o funcionamiento puede tener consecuencias perjudiciales sobre las personas. En este sentido, se trata también de hacer que los resultados sean más **previsibles**, menos inciertos y más **controlables** por el conjunto de la ciudadanía.

La ejecución de una auditoría algorítmica requiere considerar de forma previa aquellos factores que permitan establecer una ruta de trabajo, así como las fases y pasos a seguir para poder realizarla de una forma adecuada.

## 3.2 PRINCIPIOS RECTORES DE LA AUDITORÍA

La metodología de auditoría algorítmica presentada en esta guía se construye atendiendo a **cuatro pilares o principios rectores**, que no se relacionan de manera jerárquica, sino que se sitúan a un mismo plano de importancia, son **complementarios**, y deben ser tenidos en cuenta durante todo el proceso de auditoría:

### 3.2.1 CUMPLIMIENTO LEGAL Y ÉTICO

En primer lugar, todo algoritmo debe cumplir con lo dispuesto por las **normas jurídicas y deontológicas vigentes**. A este respecto, la auditoría de un algoritmo debe tener en cuenta cuál es el marco jurídico aplicable y cuáles son los derechos y valores implicados. En el caso de la protección de datos personales, como se ha explicado, es el marco establecido por el Reglamento General de Protección de Datos del Parlamento Europeo y del Consejo, así como por la Ley Orgánica 3/2018, de 5 de diciembre, de

Protección de Datos Personales y garantía de los derechos digitales, y aquellos textos jurídicos y normas sectoriales relacionados con el ámbito de actuación concreto que sea de aplicación en el caso del algoritmo auditado. Además de esto, se debe cumplir con las normas y códigos deontológicos relacionados y debe ser diseñado, implementado y revisado desde una perspectiva ética, respetuosa con las normas sociales en materia de privacidad, protección de datos, igualdad, cohesión social, libertad y confianza. Finalmente, se espera que el algoritmo respete y promueva el respeto de los derechos fundamentales que puedan verse afectados durante su diseño e implementación, más allá del derecho a la privacidad y la protección de datos (Arts. 7 y 8 CEDH). Esto incluye derechos como la integridad (Art. 3, CEDH) y libertad (Art. 5 CEDH) de las personas implicadas.

### 3.2.2 DESEABILIDAD

El segundo es relativo a la **deseabilidad del sistema**. Un algoritmo de impacto social debe ser siempre explicable, preciso, replicable, transparente y justo. Por este motivo, es imprescindible prestar atención a cuál es el "problema" al que pretende dar solución el sistema auditado, y examinar si la tecnología utilizada es, efectivamente, la mejor manera de abordarlo. La perspectiva del análisis político y cultural es imprescindible de cara a realizar un pronóstico adecuado del sistema auditado desde un punto de vista técnico y sociológico. Esto debe contribuir a que las soluciones aportadas sean lo menos invasivas posible, al mismo tiempo que cumplen de la forma más eficiente posible las expectativas y necesidades de los actores involucrados.

Que un algoritmo sea deseable implica que este no incurra en formas de discriminación a individuos o grupos y, especialmente, que no impacte de manera perjudicial sobre individuos o grupos vulnerables de manera diferente a otros individuos o grupos, reforzando o afectando de algún modo a aquellos factores que provocan su vulnerabilidad. Igualmente, importante es que el sistema no esté sesgado. En este sentido, los responsables a cargo del diseño e implementación de un algoritmo deben

contemplar aquellos factores que puedan impactar de forma general (en el ámbito de aplicación del algoritmo) y de manera diferencial (entre distintos grupos de población) en el modo en que estas personas utilizan, comprenden e interpretan las funciones, características y objetivos del procesamiento.

### 3.2.3 ACEPTABILIDAD

Un tercer aspecto crucial a la hora de evaluar un sistema algorítmico es su **aceptabilidad social**. Una auditoría algorítmica de impacto social debe preguntarse si el sistema auditado es o no aceptable desde el punto de vista social, y a ojos de la sociedad. Un sistema que tiene efectos sobre la vida de las personas, ya sea de forma directa o indirecta, debe ser comprensible, controlable, sostenible y, en alguna medida, beneficioso para las partes afectadas por el mismo. Por ejemplo, un algoritmo que clasifica los perfiles de personas solicitantes de ayuda a los servicios sociales, puede ser objeto de rechazo social o ser percibido de manera inadecuada por el público. Esto podría suceder si no se comunica de manera adecuada y transparente su funcionamiento, objetivos y resultados esperados, o bien si estos no son proporcionados o necesarios a ojos de la población. En este sentido, el artículo 13, establece la obligación al responsable de la recopilación y el procesamiento de los datos, de informar de la existencia de decisiones automatizadas, incluida la elaboración de perfiles, y de proporcionar a la persona afectada información significativa sobre la lógica de tratamiento aplicada, así como el calado y las consecuencias previstas de dicho tratamiento.

En esta línea, la información aportada sobre el algoritmo debe ser clara y suficiente como para que la ciudadanía y los clientes puedan comprender y valorar los beneficios que aporta y perjuicios que causa, así como para participar de una manera significativa, ya sea de forma directa, o a través de sus representantes públicos o perfiles especializados, en su desarrollo e implementación. Asimismo, la

aceptabilidad de un determinado algoritmo depende de que este esté bien alineado con los objetivos públicos o privados comunicados de forma explícita a los usuarios o partes interesadas. En este sentido, el diseño del algoritmo debe prestar especial atención a aquellos aspectos que puedan confrontar con los valores o características culturales predominantes en su ámbito social de aplicación. Un ejemplo sería un sistema de detección facial entrenado basado en datos de rasgos caucásicos, que no es capaz de identificar correctamente a las personas con rasgos asiáticos. No tener estos elementos en cuenta puede afectar tanto la eficiencia del sistema automático como la reputación de la organización a cargo de su diseño e implementación, al provocar posibles efectos discriminatorios.

### 3.2.4 PROTECCIÓN Y GESTIÓN ADECUADA DE LOS DATOS

En cuarto lugar, pero no por tener una menor importancia, es imprescindible que una auditoría algorítmica atestigüe que se ha hecho una **gestión responsable y adecuada de los datos** implicados a lo largo del ciclo de vida del algoritmo. Esta debe responder a los principios del tratamiento antes referidos, que establecen el RGPD y la LOPDGDD, como la exactitud, la limitación del plazo de conservación, la limitación de la finalidad o la integridad y confidencialidad de los datos.

Ello supone tener en cuenta que los datos sean de calidad, estén actualizados, procedan de fuentes fiables, sean proporcionales al objetivo perseguido por el sistema, y sean almacenados y tratados mediante técnicas pertinentes y durante un periodo de tiempo claro y preestablecido. En todo caso, los datos deben poder ser eliminados y actualizados y deben cumplir, si es preciso, criterios de anonimización adaptados a las especificidades del caso.

Debe tenerse en cuenta que la buena calidad de los datos, así como su gestión, incluida la documentación de todos los procesos de datos que afectan el entrenamiento de un algoritmo, son fundamentales no solo

para su correcto funcionamiento sino también para su transparencia hacia la persona interesada, en particular, y hacia el conjunto de la sociedad, en general. Un escaso conocimiento de los datos de entrada o salida de un algoritmo, lo puede transformar en una caja negra difícil de explicar y auditar.

## 3.3 FASES DE LA AUDITORÍA

Una auditoría algorítmica es un **proceso dinámico**, que se define de forma paralela al desarrollo y el funcionamiento del algoritmo. Por este motivo, no debe ser considerada como un conjunto inmutable de pasos, reproducibles de igual manera para cada auditoría algorítmica, sino que **deberá adaptarse al caso del algoritmo concreto y al contexto específico** de cada una de las situaciones en las que este se inscribe.

No obstante, sí es posible determinar **una serie de etapas generales que toda auditoría debería seguir**, con unos objetivos definidos. Nótese que no se está indicando aquí que estas cinco etapas deban realizarse estrictamente en el orden que se presenta a continuación. Dado el dinamismo de la auditoría, al que ya se ha hecho referencia, el proceso tiene un carácter necesariamente cíclico, que requiere ir completando la información que se detalla en cada una de las etapas, de manera que existan retroalimentaciones, aunque respetando en la medida de lo posible el orden que aquí se explica. Por otra parte, cabe destacar que la aproximación metodológica al proceso de auditoría que propone esta Guía combina diferentes técnicas de análisis cuantitativo y análisis cualitativo.

| Estudio preliminar | Mapeo | Plan de análisis | Análisis | Informe de auditoría |
|---|---|---|---|---|
| Partes implicadas<br>Problema nuevo/conocido<br>Intercambio de información<br>Diario de Auditoría. | Grado de desarrollo del sistema<br>Lista de requisitos mínimos<br>Expectativas y principales cuestiones a analizar. | Definición de los términos y plazos de la auditoría<br>Elección de metodología y equipo auditor<br>Consenso del Plan de análisis. | Investigación<br>Ejecución, seguimiento y reajustes del Plan de análisis (si corresponde)<br>Obtención y análisis de resultados. | Interpretación de resultados<br>Conclusiones y valoración final del sistema<br>Recomendaciones de mejora. |

Este apartado explica las cinco etapas del proceso de auditoría que propone la metodología presentada en esta Guía:

### 3.3.1 ESTUDIO PRELIMINAR (PUNTO DE PARTIDA): ¿QUIÉN, QUÉ Y CÓMO SE HACÍA PREVIAMENTE?

El primer paso para auditar un sistema algorítmico es comprender quién lo encarga, lo diseña, lo desarrolla, lo financia y lo implementa[5] y cuál es el problema que se pretende resolver con el uso de este algoritmo. En este punto se puede ya observar si la implementación del uso de este algoritmo implica la recopilación o el tratamiento de datos personales, en cuyo caso quedaría enmarcado dentro del ámbito del RGPD y la LOPDGDD.

Para valorar la eficiencia y la conveniencia del algoritmo, será especialmente útil comprender si quien diseña/implementa el algoritmo lo hace para abordar un problema "nuevo"[6], o si es un problema "conocido"[7], que antes se afrontaba mediante un método que no implicaba el uso de un algoritmo. Esto puede conllevar un cambio en la forma de recoger y/o procesar los datos, una variación en los datos que se recogen y/o se procesan, o bien puede suponer que comienzan a recopilarse y/o a tratarse datos, cuando antes no se hacía. En cualquier caso, habrá que hacer un estudio del riesgo que se introduce en el tratamiento por el hecho de procesar los datos mediante un sistema algorítmico.

La respuesta a estas cuestiones plantea dos escenarios diferentes para el análisis del sistema. En el primer caso (problema nuevo), será relevante comprender **cuándo y por qué** se ha tomado la decisión o se ha detectado la **necesidad de utilizar un algoritmo**. En el segundo caso (problema conocido), se tratará de acreditar **desde cuándo y por qué se aborda** este problema al que ahora se dedica el algoritmo.

Para esto es imprescindible establecer un **intercambio fluido de información** con el cliente y el equipo desarrollador del algoritmo, que permita resolver estas dudas y las que se presentan en las fases posteriores. Este intercambio de información inicial puede ser más o menos formal, dependiendo de las circunstancias y es recomendable que se realice, en cumplimiento de los principios de **responsabilidad y rendición de cuentas**, **a través de algún medio del cual quede constancia**, preferiblemente por escrito. Esto es así porque esta será una información importante para el desarrollo del resto del proceso y será útil poder volver a consultarla en fases posteriores. En esta instancia es recomendable firmar un acuerdo de confidencialidad entre el auditor y el auditado detallando los objetivos del intercambio de datos, sus medios y sus requerimientos.

---

[5] Véanse los roles de responsable y encargado establecidos por los Artículos 4, 24, 26 y 28 del RGPD, mencionados en el apartado II de esta Guía.

[6] Hablar de un problema "nuevo" no implica que el problema no existiese o no hubiera sido detectado anteriormente, sino que las personas u organización/organizaciones que diseñan, desarrollan e implementan un algoritmo no lo habían abordado antes.

[7] Se puede considerar que un problema es conocido si se ha tratado antes con este problema en concreto, o bien si se ha tratado con uno significativamente similar a partir de observaciones objetivas. Es decir, cabe la posibilidad de que se hayan utilizado algoritmos previos con el mismo fin o protocolos humanos que buscan ser reproducidos mediante el algoritmo.

Para esto es imprescindible establecer un **intercambio fluido de información** con el cliente y el equipo desarrollador del algoritmo, que permita resolver estas dudas y las que se presentan en las fases posteriores. Este intercambio de información inicial puede ser más o menos formal, dependiendo de las circunstancias y es recomendable que se realice, en cumplimiento de los principios de **responsabilidad y rendición de cuentas**, **a través de algún medio del cual quede constancia**, preferiblemente por escrito. Esto es así porque esta será una información importante para el desarrollo del resto del proceso y será útil poder volver a consultarla en fases posteriores. En esta instancia es recomendable firmar un acuerdo de confidencialidad entre el auditor y el auditado detallando los objetivos del intercambio de datos, sus medios y sus requerimientos.

Coincidiendo con el inicio de esta fase, y con el objetivo de mejorar la transparencia, la trazabilidad y la calidad del proceso, se recomienda comenzar un Diario de Auditoría que recoja información relevante sobre interacciones e intercambios de información con el cliente, decisiones importantes que se han tomado, problemas detectados, sugerencias de mejora para el momento presente o futuro, etc. Este diario se concibe, en principio, un documento interno, que podrá ser actualizado por el equipo auditor a lo largo de todo el proceso de auditoría, para recoger aquella información imprescindible.

### 3.3.2 MAPEO DE LA SITUACIÓN: ¿CÓMO, CUÁNDO, POR QUÉ Y PARA QUÉ DESARROLLA E IMPLEMENTA QUÉ ALGORITMO? ¿CUMPLE UNOS REQUISITOS MÍNIMOS PARA SER AUDITADO?

Esta segunda etapa se dedica a recopilar **información básica** sobre el algoritmo y el contexto en el que se inscribe y al que afecta. Tiene dos objetivos principales: El primero es averiguar si se cumplen o no unos requisitos mínimos que permitan **determinar si el algoritmo puede ser auditado** con garantías de calidad. Para ello, en las próximas páginas se establece un listado de requisitos a cumplir para la realización de la auditoría. El segundo es **esclarecer las expectativas del estudio e identificar las principales cuestiones a analizar** en la auditoría. Esto permitirá, a su vez, elaborar el Plan de análisis, que constituye la siguiente fase (3) de la auditoría.

Una primera cuestión a la que atender es el **grado de desarrollo del algoritmo**. Es decir, el algoritmo a auditar puede ser un proyecto que todavía no ha comenzado o que está en una etapa incipiente, puede estar en diseño o desarrollo; puede estar ya diseñado, evaluado o entrenado; puede estar en fase de funcionamiento (esto puede implicar que esté en interacción con el mundo), o puede ser algoritmo que ya ha sido utilizado.

Tener claro el grado de desarrollo del algoritmo desde el inicio es importante, porque dependiendo de dicho grado de desarrollo del algoritmo, el proceso de auditoría variará, dado que no en todas las fases se dispone de la misma información, ni es posible el mismo tipo de medidas de corrección, reelaboración o mitigación de sesgos, por ejemplo. Volveremos más adelante sobre esta cuestión.

Uno objetivo principal de la segunda fase de auditoría es obtener información básica sobre el algoritmo, que permita comprobar si es posible auditarlo. Así, en este momento será oportuno enmarcar el problema del algoritmo en el ámbito del Registro de Actividades del Tratamiento (RAT) asociado al caso, en cumplimiento del Artículo 30 del

RGPD. De acuerdo con esto, cada responsable y, en su caso, el encargado del tratamiento de datos personales, llevarán un registro de las actividades de tratamiento efectuadas bajo su responsabilidad.

A continuación se expone una **lista de requisitos** que el cliente de la auditoría debe comprometerse a cumplir, para que algoritmo pueda ser auditado con garantías de calidad[8].

[8] Al igual que el resto de la metodología presentada en esta Guía, este es un listado de requisitos original, elaborado según los puntos que deben figurar en el Registro de Actividades del Tratamiento (ver la nota anterior de este documento y el Art. 30, RGPD), la experiencia auditora del equipo de investigación de Eticas Research and Consulting y a textos académicos previos, entre los que destaca el trabajo de Mitchell, *et al.* (2019).

### *3.3.2.1 LISTA DE REQUISITOS DE RECOMENDADO CUMPLIMIENTO PARA QUE UN ALGORITMO PUEDA SER AUDITADO CON GARANTÍAS DE CALIDAD:*

- **Datos identificativos y de contacto** de la/s persona/s o institución/instituciones encargadas y responsables de los distintos aspectos relativos al diseño, el desarrollo y la implementación del sistema y, en su caso, del corresponsable, del representante del responsable, y del delegado de protección de datos;

- **Fecha de creación** del algoritmo y, a ser posible, la versión del mismo[9].

- **Licencia** del algoritmo. En este registro cabe tener en cuenta si la propiedad del algoritmo es pública o privada y las condiciones contractuales existentes entre el desarrollador y el responsable del uso del algoritmo. Esto puede ser un elemento que limite el acceso al código del algoritmo.

- Datos sobre la **arquitectura básica** del algoritmo, incluyendo datos sobre la forma de aprendizaje, entrenamiento y funcionamiento del sistema.

- **Otros detalles de referencia y especificaciones sobre el algoritmo**, no reflejadas en los apartados anteriores como: artículos o publicaciones que contengan más información sobre el algoritmo, datos de citación del algoritmo, o datos de retroalimentación de su funcionamiento.

[9] Si este es un algoritmo desarrollado a partir de versiones anteriores del mismo algoritmo, será útil saber en qué difiere esta versión de las anteriores.

- **Marco teórico** sobre el cual se desarrolla el modelo.[10]

- **Marco metodológico y explicación de la metodología utilizada para definir el modelo (incluyendo las asunciones de base).**

- **Acceso e información sobre el código del algoritmo: Este debe respetar unas normas de calidad. Esto quiere decir que, la información sobre el código, incluirá aquella información y aclaraciones sobre el mismo necesarias para su inteligibilidad, como: el/los lenguaje/s de programación; notas aclaratorias; programas, paquetes y librerías necesarios para su lectura, etc.**

- **Acceso a información sobre la API** del algoritmo (interfaz de programación de aplicaciones, por sus siglas en inglés), si se ha desarrollado.

- **Acceso a información sobre la/s base/s de datos** utilizada/s para el desarrollo del algoritmo, y a las bases de datos utilizadas para su entrenamiento (*base de datos de entrenamiento*) y su testeo o evaluación (*base de datos de testeo*). A este respecto, el cliente debe aportar información sobre, al menos, la/s fuentes de la/s que se obtienen los datos recogidos en las bases de datos y las motivaciones por las que se han elegido estos datos y las categorías de datos utilizados (no personales, personales, sensibles…).

[10] El equipo auditor podrá consultar esta información antes o después de la elaboración del Plan de análisis, en función de cómo considere que esto puede condicionar la objetividad de la auditoría.

Del mismo modo que el código, las bases de datos que nutren el algoritmo deben de respetar unas normas de calidad que las hagan legibles, comprensibles y usables. Por ello:

- las bases de datos deben tener una estructura ordenada y coherente entre ellas;
- en la medida de lo posible, los datos deben ser calidad, exactos y actualizados, es decir, contener el mínimo número posible de registros inválidos;
- las variables y la cantidad de datos asociados a las mismas deben ser claramente identificables y manejables;
- indicar si se han realizado operaciones de anonimización o pseudonimización de los datos;
- se recomienda que las bases de datos vengan acompañadas de un diccionario que permita su mejor comprensión.

- Definición de las categorías de interesados afectados por la implementación del sistema y/o cuyos datos son objeto de tratamiento por el mismo, incluyendo los **grupos involucrados** en el algoritmo y la descripción de sus variables identificativas, especialmente, de aquellos considerados **grupos vulnerables**, bien por quien desarrolla e implementa el algoritmo, o bien por el equipo auditor. En su caso, también se recomendará identificar, como parte de este punto, **organizaciones de carácter social**, cuya labor se centra en la mejora de las condiciones de vida de estas personas o grupos vulnerables.

- Información sobre el **entrenamiento y la evaluación** del modelo, incluyendo:
  - frecuencia y distribución de datos y variables en la/s base/s de datos;
  - información sobre el pre-procesamiento de los datos, su procesamiento durante el desarrollo del modelo y su post-procesamiento;

- parámetros y criterios aplicados para la conseguir la imparcialidad del modelo, o que sirvan para la evaluación interna de su efectividad;
- cuando sea posible, una descripción general de las medidas técnicas y organizativas de seguridad implementadas en el algoritmo.

- **Finalidades o usos previstos** del algoritmo: nociones iniciales sobre quién, cómo lo utiliza y para qué lo hace. Incluyendo:
  - Usos y fines principales del uso del algoritmo;
  - Usuarios principales del algoritmo;
  - Posibles usos y usuarios secundarios[11]
    - Categorías de destinatarios a quienes se comunicaron o comunicarán los datos personales que trata el algoritmo, incluidos los destinatarios en terceros países u organizaciones internacionales;
  - En su caso, las transferencias de datos personales a un tercer país o una organización internacional, incluida la identificación de dicho tercer país u organización internacional y, en el caso de las transferencias indicadas en el Artículo 49.1, párrafo segundo, la documentación de las garantías adecuadas a este respecto;
  - Cuando sea posible, los plazos previstos para la supresión de las diferentes categorías de datos;

- **Objetivos** del uso del algoritmo: qué pretende conseguirse con el uso del algoritmo, en términos cuantitativos y cualitativos.

[11] Nótese que esta es una cuestión particularmente sensible cuando se trata de la protección de los datos personales, dado que utilizar datos que han sido recogidos para un determinado fin, para otro diferente, revela una mala práctica que podría pasar desapercibida. Estos, por lo tanto, deberían estar informados y tener una base de legitimación.

En el caso de abordar un *problema nuevo*, se requerirá al cliente explicación sobre las motivaciones y argumentos que le mueven a abordar este problema. En el caso de abordar un *problema conocido*, el cliente aportaría información sobre si los objetivos son los mismos que se perseguían mediante el esquema de funcionamiento anterior, o si han cambiado. Valorar el algoritmo en función no solo de su uso, sino de sus objetivos, permitirá realizar una evaluación más ajustada del sistema.

- Información sobre **las dinámicas, actividades y procesos** en los que se integra el sistema. Esto incluye detalles sobre el equipo que trabaja con el sistema, los procesos organizacionales en los que se integra y las actividades y las dinámicas internas de las que forma parte, o que se modifican con su introducción. Del mismo modo que en el caso de los objetivos, será importante saber si estas cuestiones se mantienen más o menos invariables con respecto al esquema de funcionamiento anterior (problema conocido).

- Información sobre las **responsabilidades de las partes implicadas** con respecto al funcionamiento del modelo. Esto incluye profundizar en la distribución de las responsabilidades de los desarrolladores en relación con el sistema y las responsabilidades de los impulsores del sistema en relación con su funcionamiento. Dicho esquema de competencias está especialmente relacionado con el concepto de rol relativo al tratamiento de los datos y a la distribución de responsabilidades de cada encargado del tratamiento, como se desarrolla en el Capítulo 4 del RGPD. En quién recaen la responsabilidad y la rendición de cuentas dependerá de quién haya diseñado, desarrollado, encargado e implementado el sistema. Por ejemplo, los desarrolladores pueden ser los encargados del tratamiento si el sistema no se desarrolla dentro de la misma organización que lo implementa. Por ello, es imprescindible delimitar responsabilidades y funciones dentro del desarrollo del algoritmo y la solución completa.

- Información sobre **factores condicionantes** de la efectividad del sistema, como son: el contexto socioeconómico/medioambiental, los instrumentos utilizados para capturar los datos de entrada del modelo, los recursos disponibles, las políticas y normas aplicables, u otros factores que puedan variar el funcionamiento del sistema. En el caso de que el problema a abordar por el algoritmo sea conocido, será valioso conocer si estas circunstancias que acompañan a la resolución del problema son las mismas que antes de la implementación del algoritmo.

Cabe notar que este es un listado de requisitos, de carácter orientativo y no exhaustivo, de la información mínima que debería estar a disposición del equipo auditor, para poder evaluar un algoritmo. También se debe considerar que, dependiendo del grado de desarrollo del algoritmo, o su tipología, puede resultar imposible proporcionar parte de esta información, en cuyo caso, deberá ser proporcionada más adelante y el cliente debe comprometerse a hacerlo cuando esté disponible.

El hecho de que **el cliente no pueda o no tenga la voluntad de proporcionar alguna de esta información**, mermará la calidad de la auditoría, poniendo en riesgo tanto su realización como su garantía de calidad. No obstante, si esto no se produce de manera recurrente para varios requisitos, o para alguno especialmente importante, no implicará necesariamente que la auditoría no pueda realizarse, sino que deberá ser el equipo auditor el que, desde su conocimiento y experiencia, valore el impacto valore el impacto de la falta de estos requisitos en la calidad de la auditoría y determine si esta debe continuar adelante o no.

### *3.3.2.2 ALGORITMO "AUDITABLE"*

Como se ha indicado en apartados anteriores de esta Guía, todo algoritmo debe poder ser auditado. No obstante, esta guía se centra en aquellos algoritmos que pueden tener un impacto social, especialmente ligado al incumplimiento de los derechos fundamentales a la protección de los datos personales y a la privacidad.

Desde la perspectiva de esta Guía de Auditoría Algorítmica, **un algoritmo debe ser auditado siempre que** este recopile o trate datos personales o sensibles, pueda afectar a la vida de las personas y/o a grupos sociales relevantes o grupos vulnerables (especialmente si afectan a cuestiones como el acceso a la educación, al trabajo, a prestaciones o beneficios sociales y o que funcionen en ámbitos como el judicial o el de la salud pública), pueda incurrir en alguna de las formas de impacto social referidas, o que pueda implicar alguna forma de discriminación o sesgo en alguna fase de su ciclo de vida.

Además, un algoritmo **podrá ser auditado siempre que cumpla con un mínimo de requisitos** de los expuestos en el listado de requisitos recogidos en este apartado, según el criterio del equipo auditor. En este sentido, las razones por las que un algoritmo es o no auditable tiene que ver con una diversidad de cuestiones, tanto relativas al mismo algoritmo, como al contexto en el que se inscribe, las personas responsables del mismo, o cuestiones administrativas y legales.

En este punto, se recomienda también tener una **primera toma de contacto con todas las partes implicadas** en el proceso de desarrollo e implementación del algoritmo y también con las partes afectadas por el mismo (personas y grupos interesados, incluyendo organizaciones de carácter social como se ha mencionado antes). Para ello, se puede

continuar con las dinámicas de intercambio de información establecidas en el primer paso de la auditoría, o bien se pueden realizar entrevistas, grupos de discusión o encuestas breves.

### 3.3.3 PLAN DE ANÁLISIS: ¿CÓMO, CUÁNDO Y PARA QUÉ SE DESARROLLA LA AUDITORÍA?

Una vez comprobado que el sistema cumple los requisitos mínimos para ser auditado, el siguiente punto es definir el **Plan de análisis de la auditoría y consensuarlo** con el cliente. Esto consiste, principalmente, en identificar, definir y consensuar con el cliente el **objeto de estudio** de la auditoría, sus **objetivos específicos, hipótesis y preguntas de investigación**, la **metodología y las técnicas** de análisis, los **parámetros de interpretación** de los resultados o los **plazos orientativos** de la auditoría.

Del mismo modo, basándose en la información de la que se dispone hasta este punto, en esta fase del análisis se deberá **definir la composición del equipo auditor** adecuado para estudiar el caso concreto, y que vendrá condicionado por factores como el tipo de sistema que se utilizará o del sector en el que se inscribe el modelo. Este equipo, tanto en el caso de las auditorías internas, como de las externas, pueden incluir personal de la entidad auditora y de la auditada, aunque se recomienda que se asegure la independencia de quienes trabajan en el análisis del modelo, para su mayor objetividad. El equipo auditor debe incluir **perfiles técnicos** como analistas capaces de realizar la parte técnica de la auditoría, especialmente los científicos de datos, y **perfiles sociales**, como los sociólogos o los expertos legales, capaces de sacar a la luz las implicaciones socio-económicas, jurídicas y éticas más profundas de los sistemas.

Por otra parte, para definir correctamente el Plan de análisis, es recomendable **revisar el marco teórico y metodológico** sobre el cual el

cliente ha desarrollado el sistema, así como esquematizar las primeras nociones sobre las **actividades y procesos** de la organización en las que éste se integra. También es importante realizar un **estudio sobre el contexto jurídico, social y económico** concreto en el cual se va a implementar el sistema, que permita comprender mejor la aceptabilidad y la deseabilidad del sistema en su conjunto. Esta información será completada durante el proceso de auditoría, y se incluirá en el Informe de auditoría.

Como parte de la elaboración del Plan de análisis, corresponde al equipo auditor **delimitar la metodología prevista** para la auditoría que, como se ha mencionado antes, será variable en función de cada caso concreto. Esto incluye concretar, al menos, los siguientes aspectos, de manera consensuada con el cliente:

- Las **partes del sistema** a auditar.

- Las **variables** principales o "variables madre" sobre las que se va a realizar el estudio de auditoría.

- Las **intersecciones** entre variables a estudiar (si corresponde).

- Los **grupos** a monitorizar y sus variables definitorias.

- Los **métodos, métricas y técnicas de análisis** cuantitativo (estadísticos, cuestionarios…) respectivos. En la próxima sección se ofrecen ejemplos.

- Los **métodos y técnicas de análisis cualitativo** del sistema (grupos de discusión, entrevistas, observación participante/no-participante,

análisis etnográfico, etc.) y las personas o grupos a las que se planteará participar en el estudio.

- Los **parámetros de interpretación** de los resultados, que deberán establecerse de acuerdo con el cliente. Esto incluye:
  - aquellos porcentajes de muestra de las distintas variables, que se consideran representativos dentro de las bases de datos;
  - aquellos porcentajes y cifras de corte significativos para la interpretación de las mediciones realizadas;
  - las medidas mínimas o máximas de precisión, deseabilidad y aceptabilidad del sistema, si corresponde.

- Los **pasos a seguir** en la auditoría.

- Los **plazos estimados** de ejecución del Plan de análisis.

- El **calendario orientativo de reuniones** de seguimiento.

Una vez definido el Plan de análisis por el equipo auditor, que detalla asimismo los entregables y el calendario de trabajo, deberá compartirse con el cliente y ser consensuado por ambas partes, antes de proceder a su ejecución. En caso de discrepancias, se podrán reajustar los términos del análisis para que ambas partes estén conformes.

Llegado este punto, el equipo auditor podrá realizar una serie de **recomendaciones preliminares** para la mejora del sistema, conforme a lo observado hasta el momento. Cabe tener en cuenta que dichas recomendaciones tendrán diferente alcance dependiendo del grado de desarrollo del algoritmo, pudiendo contribuir desde el inicio a su remodelado en el caso de los sistemas con avanzado grado de desarrollo o simplemente sugiriendo consideraciones generales para su implementación en el caso de los que se encuentren en su fase inicial. Este es un punto importante para la auditoría, que, como hemos explicado

es un proceso cíclico, dado que si en este punto ya se han detectado problemas sustanciales o cuestiones importantes a las que atender, puede que la auditoría requiera ver cómo el cliente responde a estos requerimientos, para poder continuar con el proceso.

Igualmente, el equipo auditor estará en situación de **identificar dificultades u obstáculos** para continuar con la auditoría. En caso de que estos sean tales como para que dicho equipo considere que no es posible continuar, la auditoría se podrá detener temporalmente, a la espera de que estos sean subsanados (con el consiguiente aplazamiento de los pasos siguientes), o bien se podrá desestimar la posibilidad de prolongar el estudio, acompañando dicha decisión de un informe razonado y argumentado de los motivos y resultados obtenidos hasta el momento. En relación con las observaciones realizadas en este punto, podrá ser necesario una reelaboración o un reajuste del Plan de análisis, que será nuevamente consensuado por las partes implicadas.

### 3.3.4 ANÁLISIS: EJECUCIÓN DEL PLAN DE ANÁLISIS

Esta fase consiste en la **ejecución del Plan de análisis** definido y acordado con el cliente. Nótese que, durante el proceso de análisis podrá surgir la necesidad de reajustar aspectos relativos a la metodología, los plazos y los objetivos de la auditoría, que se irán consensuando con el cliente. El estudio del algoritmo, como se ha mencionado antes, se compone de **dos partes más o menos diferenciadas**, que se corresponden con el análisis del sistema desde la perspectiva **cuantitativa** y la **cualitativa**.

En primer lugar, antes de proceder a los análisis planificados, el equipo auditor deberá **completar la revisión del estado de la cuestión** relativo a los aspectos detallados en el Plan de análisis, para poder analizar adecuadamente los resultados obtenidos. Es decir, en este punto el equipo auditor realizara una revisión de las **teorías de base** para la

creación del modelo, de los **razonamientos existentes tras asunciones importantes** para su desarrollo (por ejemplo, examinar cuáles son los argumentos tras una relación causal que modela un algoritmo, como la selección de variables que definen un fenómeno), y de las **metodologías** utilizadas. Asimismo, realizará un estudio de aquellos **aspectos relativos al contexto** en el que se diseña, se desarrolla y se implementa el algoritmo, ya sean aspectos sociales, económicos, organizacionales, medioambientales, técnicos, científicos, o de cualquier otra clase. Esto consiste básicamente, en **conocer lo mejor posible la realidad en la que se integra el sistema**, para analizar sus posibles implicaciones en su contexto real.

#### *3.3.4.1 LA AUDITORÍA TÉCNICA*

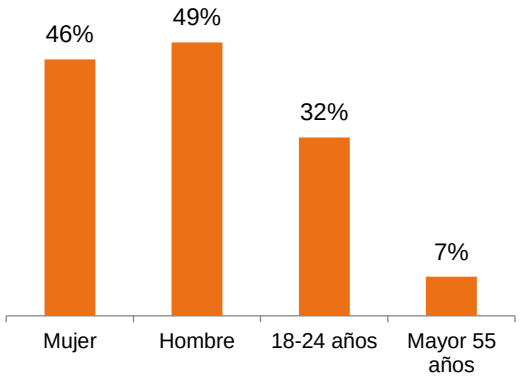
Comenzando por las orientaciones relativas a la **auditoría de carácter cuantitativo** del algoritmo, se realizará una **descripción de la/s bases de dato/s** utilizadas por el cliente para desarrollar, entrenar y evaluar el sistema. Asimismo, se examinará la validez de las muestras relativas a las variables y los grupos relevantes para el estudio. Con este fin, se hará un primer trabajo descriptivo de **identificación, cuantificación y análisis de la frecuencia y la distribución de las variables, intersecciones entre variables y grupos** relevantes para el estudio en la base de datos (incluyendo los grupos protegidos). Para ello se tendrá en cuenta la información proporcionada por el cliente de la auditoría acerca de cuáles son las variables que han considerado más relevantes para el desarrollo del modelo.

También se estudiará si el sistema trabaja con variables *proxy*, especialmente si estas variables proxy son relevantes para el algoritmo. Las variables proxy son aquellas variables que no tienen un gran interés cuando están aisladas, pero pueden revelar (mediante inferencias) información importante (o sensible) cuando se analizan en conjunto con otras variables. Por ejemplo: si el algoritmo se dedica a predecir la

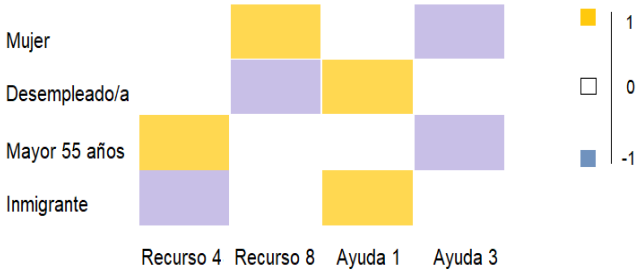
posibilidad de sufrir una vulneración de derechos, y se define que una persona está en riesgo de sufrir una vulneración de derechos a partir del análisis de las variables X, Y y Z, entonces “estar en riesgo de sufrir una vulneración de derechos” no sería una variable recogida de forma explícita en la base de datos, sino que se llegaría a ella a partir del análisis de estas otras tres variables proxy (X, Y y Z). Analizar la solidez de estas relaciones es de interés para la auditoría, dado que estas pueden variar el modelo de manera determinante.

A continuación, se muestran unos gráficos básicos a modo de ejemplo, que representan los resultados ficticios de un análisis de frecuencia de variables y correlaciones positivas y negativas más fuertes entre variables:

Ej. Gráfico de frecuencia de variables sociodemográficas[12]




*Ej. Gráfico de correlaciones entre variables sociodemográficas y asignación de recursos/ayudas*[13]




12 y 20 Estos dos gráficos simplificados representan los resultados ficticios de un análisis descriptivo de la base de datos utilizada para entrenar un algoritmo de adjudicación de recursos:

El primero de ellos recoge el número de personas de la base de datos que son mujeres, hombres, o tienen entre 18 y 24 años o más de 55 años. Vemos que la frecuencia del grupo mayores de 55 años es notablemente más baja que el resto que de grupos de edad.

El segundo gráfico muestra que es más probable que este modelo adjudique a una mujer el recurso 8 y menos probable que le asigne la ayuda 3; es más probable que asigne a una persona desempleada la ayuda 1 y menos el recurso 8; es más probable que asigne el recurso 4 a una persona mayor de 55 años y menos que le asigne la ayuda 3; y que es más probable que asigne la ayuda 1 a una persona inmigrante, y menos que le asigne el recurso 4. El gráfico también muestra que el modelo no muestra correlaciones significativas para la adjudicación de determinadas ayudas o recursos a ciertos grupos. [13]

El análisis de los resultados de estas primeras mediciones permitirá obtener unas conclusiones iniciales acerca del algoritmo auditado, emitir recomendaciones preliminares sobre la base del análisis de los datos, si corresponde y plantear preguntas de investigación o hipótesis fundamentadas, acerca del funcionamiento del sistema.

Por poner un ejemplo: si se observa una distribución muy desigual entre variables, correlaciones espurias entre ellas, o una utilización inadecuada de variables proxy, se podrá plantear una pregunta de investigación, o bien hipotetizar, según la teoría al respecto, que el modelo podría tener un comportamiento desventajoso para un grupo vulnerable en cuya definición se incluyen estas variables. Estas hipótesis se irán refutando a lo largo del estudio y se plasmarán en el Informe de auditoría.

En este contexto, se valorará si **la muestra de las variables de estudio es suficiente** en la base de datos. Para determinar qué variables y/o grupos se puede analizar considerando que son modelables de forma robusta en el sistema y cuáles no, se recomienda tomar como referencia que estas tengan, al menos, más de un 5% de representatividad en la muestra. Por debajo de este porcentaje, la variable y/o grupo puede tener una representación demasiado escasa en la base de datos, lo cual, no obstante, debe ser señalado en el estudio, dado que podría estar afectando a la precisión del modelo. Sin embargo, como se ha indicado en la sección anterior, estos porcentajes de interpretación serán en todo caso acordados con el cliente durante la elaboración del Plan de análisis.

En caso de que alguna de las variables no alcance el mínimo muestral acordado en el Plan de análisis, se podrán hacer recomendaciones a este respecto, como solicitar que se revise la forma de recoger información relativa a estas variables o las cantidades de datos al respecto, o bien se podrá variar la relación de variables y/o grupos a estudiar, con respecto a lo acordado en el Plan de análisis (esta dificultad y las medidas de

corrección se reflejarán en el Informe de auditoría). Al mismo tiempo, la **representatividad de una base de datos** —por ejemplo, en el caso de que esta se refiera a la composición sociodemográfica de los individuos o grupos presentes en ella—, podrá ser cuestionada en relación a los resultados de análisis de la frecuencia y distribución de variables: bien en relación al conjunto de una población dada, bien dentro de la misma base de datos, bien en relación a algún grupo concreto, u otro punto de referencia. Este punto servirá además para comprobar que las fuentes de datos del sistema son fiables y suficientes, y que se está realizando una gestión adecuada de los datos, a nivel cuantitativo.

Por otra parte, se estudiará si **la distribución de las variables es adecuada**, o si el sistema presta demasiada o muy poca atención a alguna de ellas. Si pensamos en un algoritmo utilizado para la asignación de recursos, por ejemplo, será interesante estudiar cómo se asignan estos recursos (¿cuántos, cuáles, a quién/es, cómo?). También, como parte de este mapeo inicial, se valorará cuál es la lógica tras las correlaciones más fuertes entre variables.

Una vez realizado el análisis descriptivo, se procederá a detectar y estudiar el sesgo algorítmico, implementando las **mediciones acordadas en el Plan de análisis**.

Es importante señalar que esta metodología de análisis de sesgo no sólo estudia la afectación del sistema sobre grupos protegidos fundamentales (estos se señalan en el apartado III de definiciones de esta Guía), sino que se centra en **detectar dinámicamente y analizar aquellos grupos vulnerables que puedan ser discriminados por el sistema**, dado el caso y el contexto en el que este se inscribe. Cabe tener en cuenta que un grupo vulnerable puede definirse mediante intersecciones entre variables (como por ejemplo: mujeres jubiladas, no blancas y con bajos ingresos, u otras combinaciones de variables personales y temporales que afectan a

las comunidades pertinentes). Por lo tanto, la definición de grupos vulnerables potencialmente afectados por un algoritmo debe hacerse en correspondencia con la realidad en la que se inscribe. Para ello, se deben identificar los patrones de vulnerabilidad y exclusión que existen dado un caso concreto, a través del análisis de las teorías y asunciones de base, así como de las variables, las intersecciones de variables, las combinaciones variables proxy y las funciones utilizadas por el sistema. Este caso ilustra por qué una auditoría algorítmica no debe consistir únicamente en un análisis cuantitativo, sino que también se debe realizar un análisis cualitativo, capaz de comprender el sistema en su marco de implementación.

Desde una perspectiva cuantitativa, la **metodología de análisis de sesgo** de un sistema algorítmico, se divide en **cuatro pasos** principales[14]:

### i. ASIGNACIÓN DE DATOS A GRUPOS

El primer paso consiste en definir la **asignación de datos a grupos concretos**, en función del trabajo de mapeo del algoritmo desarrollado previamente. Esto quiere decir que se clasifican los datos relativos a características o atributos determinadas en grupos, que pueden estar superpuestos (asignación "blanda") o no superpuestos (asignación "fuerte"). Esta superposición hace referencia a la convergencia de más de una característica protegida, como por ejemplo: “mujer con bajos ingresos”. En la mayoría de los casos, los grupos se realizarán según características únicas. Se puede utilizar cualquier característica asignada a múltiples individuos para crear tales grupos, pero se presta especial atención a los atributos protegidos. Estas agrupaciones se crean para

[14] Este esquema se deriva de la metodología aplicada por Carlos Castillo, investigador del Departamento de Tecnologías de la Información y las Comunicaciones de la Universitat Pompeu Fabra, en trabajos previos realizados, en colaboración con Eticas Research and Consulting.

evaluar en qué medida un algoritmo puede tratar o afectar a un grupo de manera diferente a otro.

### ii. IDENTIFICACIÓN DE GRUPOS VULNERABLES

El segundo paso consiste en determinar **qué grupos de los que se han definido se consideran grupos vulnerables o protegidos en el contexto específico de la auditoría**, lo que significa que no deben verse desfavorecidos por la aplicación del algoritmo y, por lo tanto, que se monitoreará de forma especial su impacto sobre ellos. Una definición acotada de grupo protegido podría basarse en el propósito de una tecnología y, por lo tanto, en la conveniencia del algoritmo. Por ejemplo, si la intención de un cierto algoritmo es aumentar la protección de los niños de cierta edad que sufren abuso doméstico, entonces los niños de esa edad constituyen un grupo protegido.

### iii. DEFINICIÓN DE LOS CRITERIOS Y LAS MÉTRICAS DE ANÁLISIS

El tercer paso **determina el conjunto de métricas** que se utilizarán para el análisis de estos grupos protegidos. El objetivo es analizar si el **algoritmo se comporta de forma adecuada en relación con los diferentes grupos** identificados, en función de criterios específicos de "justicia algorítmica". Existen múltiples definiciones de justicia o equidad algorítmica[15]. De entre las más comúnmente aceptadas sobresale una definición ligada a la **equidad grupal**[16], que comprende que un algoritmo **no debe producir resultados desventajosos para grupos específicos o vulnerables**.

A menudo, y de forma general, se entiende que existe equidad grupal si se cumplen una o más de las siguientes condiciones:

- La probabilidad de que un algoritmo genere un resultado no viene determinada por el atributo que define a un grupo específico (independencia);

- Esto es así aunque los datos de la realidad acompañen la asignación de un resultado a un determinado grupo (separación);

- Y la medición realizada por un algoritmo no se combina con atributos protegidos para obtener un resultado (suficiencia).

No obstante, estas condiciones no pueden cumplirse en determinados casos, en los cuales es necesario ligar los resultados a la presencia de atributos protegidos de forma explícita, para cumplir los objetivos perseguidos. Por otra parte, aunque estas definiciones de justicia algorítmica se centran en grupos y no garantizan que un algoritmo se comporte de manera equitativa con diferentes individuos, la literatura académica al respecto indica que es complejo desarrollar mecanismos consistentes para medir el tratamiento desigual a nivel individual. Esta es una forma de medición, que algunos autores consideran que podría perjudicar las medidas de equidad grupal, al obviar **factores contextuales** más amplios[17].

Por este motivo, entre otros, **el** marco **contextual en el que opera un algoritmo debe ser analizado**, tanto desde un punto de vista cuantitativo como cualitativo, y utilizado para interpretar sus resultados en términos de justicia algorítmica. Esto es especialmente importante en aquellos casos en los que un algoritmo se utiliza para ordenar elementos como personas, grupos de personas o categorías similares. En este caso, se recomienda: por una parte, que exista una **presencia suficiente de**

---

[15] Para más información véanse los trabajos de Binns *et al.* (2018), Castillo (2019), Chouldechova (2017), Dwork e Ilvento (2018), Dwork *et al.* (2012), Holstein (2019), Kim *et al*. (2018), Kleinberg *et al.* (2017), Kyung Lee (2018) o Nayanan (2018), recogidos en el apartado de Referencias de esta Guía.

[16] Para más información véase el trabajo de Barocas y Hardt (2017), recogido en el apartado de Referencias de esta Guía.

[17] Para más información véanse los trabajos de Heidari *et al.* (2018) y Speicher *et al*. (2018), recogidos en el apartado de Referencias de esta Guía.

**elementos definitorios del grupo protegido**, para poder monitorizar que el algoritmo no incurra en formas de discriminación y tratamiento diferencial a nivel grupal; y por otra, **que se traten de manera consistente** los elementos relativos a los grupos, para evitar formas de discriminación individual, es decir, que las posibles diferencias en el tratamiento de las personas vengan únicamente determinadas por sus atributos no protegidos (Castillo, 2019).

Como se ha indicado antes, existen múltiples definiciones de métricas posibles para evaluar el sesgo algorítmico, y su elección dependerá de cuestiones relacionadas con la forma de funcionamiento del algoritmo, sus objetivos, el tipo de información que maneja, entre otras cosas. Sin embargo, se debe mantener un cierto grado de acuerdo. Las métricas que se señalan aquí parten de la base de que el algoritmo **puede tener un resultado positivo o favorable u otro negativo o desfavorable**, o bien que es posible **ordenar estos resultados en una escala que va de los más positivos a los más negativos**, o viceversa (por ejemplo, un algoritmo clasifica a los solicitantes para un trabajo).

Para valorar si un sistema trata efectivamente de manera equitativa a los diferentes grupos afectados, se aconseja como proceso general, aplicable a diferentes casos, estudiar si el análisis del sistema reporta **tasas de impacto o tratamiento diferencial** y si existen diferencias significativas entre las **tasas de falsos positivos** (FNR, por sus siglas en inglés) **y las tasas de falsos negativos** (FPR, por sus siglas en inglés), entre diferentes grupos. Esto consiste, en primer lugar, en cuantificar la medida en que un algoritmo tiene un impacto diferente en diferentes personas o grupos y la medida en que trata a personas o a grupos de personas de manera diferente. En segundo lugar, se busca identificar si existen diferencias desfavorables para el grupo protegido entre las tasas de falsos positivos, falsos negativos, verdaderos positivos o verdaderos negativos, asignados a este grupo en comparación con otro. Es decir, se trata de examinar si un sistema sobreestima o subestima a un determinado grupo,

de forma relevante con respecto a otro grupo y en relación a los objetivos del sistema. En general, estas métricas cuantifican la medida en que un algoritmo trata a las personas de manera diferente (*disparate treatment, DT*) y la medida en que un algoritmo tiene un impacto diferente en diferentes personas (*disparate impact, DI*).

Para esta evaluación de sesgos, se recomienda utilizar herramientas estándar como Aequitas Bias and Fairness Audit Toolkit[18], AI Fairness 360 Open Source Toolkit[19] o Algorithmic Equity Toolikt[20], entre otras.

### iv. APLICACIÓN DE LAS MÉTRICAS AL ANÁLISIS DE GRUPOS

El cuarto paso consiste en **aplicar las métricas escogidas, relevantes para el caso concreto, y analizar sus resultados para los grupos seleccionados**. Si los datos son procesados en varias etapas en un sistema (como la recopilación de datos y el análisis de datos), el análisis de estas métricas se lleva a cabo para cada etapa o paso por separado. A modo de ejemplo se listan a continuación algunas posibles métricas y valores ficticios de análisis:

- **Ratio de impacto**, este ratio se calcula como el porcentaje del grupo protegido con predicción/resultado positivo dividido entre el porcentaje del grupo no protegido con predicción/resultado positivo. Típicamente, valores inferiores al 80% se consideran problemáticos y hay que revisar más detenidamente si dicha disparidad se debe a un caso de

[18] Para más información, consultar la página web de Aequitas: http://www.datasciencepublicpolicy.org/projects/aequitas/.
[19] Para más información, consultar la página web de AI Fairness 360: https://aif360.mybluemix.net/.
[20] Para más información, consultar la página web de Algorithmic Equity Toolkit. https://aekit.pubpub.org/.

discriminacion algorítmica. Valores cercanos al 100% se consideran más equitativos.

- **Tasas de falsos positivos y falsos negativos**. Un falso positivo es una predicción positiva en la realidad que resulta ser negativa en los resultados algorítmicos. En sentido opuesto, un falso negativo es una predicción clasificada como negativa que resulta ser positiva en el caso real. Se considera que un grupo tiene riesgo subestimado por el algoritmo si la tasa de falsos negativos es mayor que la de falsos positivos (siendo esta última mayor que 0). Por otra parte, se suele considerar que existe disparidad entre grupos si las tasas de falsos negativos asignados a grupos comparados tienen una diferencia substancial.

Supongamos que estamos analizando un algoritmo que predice el riesgo de una población en situación de pobreza para, de este modo, priorizar de forma más eficiente los recursos sociales y asignarlos a aquellas personas con riesgo alto. Estamos analizando el tratamiento diferencial por grupo y hemos comprobado que el algoritmo arroja estos resultados:

| Grupo | Tasa de falsos negativos | Tasa de falsos positivos |
|---|---|---|
| A | 0.55 | 0.14 |
| B | 0.72 | 0.12 |

En esta tabla vemos que el algoritmo asigna un riesgo alto de forma más frecuente al grupo A. Si calculamos el ratio de impacto como 30%/60% vemos que su valor es 50%, valor inferior al valor referencial de 80%. Esto implica que si el grupo A es el menos desprotegido de entre los grupos comparados, la disparidad observada en el tratamiento no supondría, en principio, discriminación.

| Grupo | Población | Predicción de alto riesgo | Porcentaje de predicción |
|---|---|---|---|
| A | 80 | 48 | = 48/80 = 60% |
| B | 40 | 12 | = 12/40 = 30% |

Aquí vemos que es más probable que el grupo B tenga una tasa de falsos negativos sustancialmente mayor que el grupo A (0.55 / 0.72 = 76%). Esto quiere decir que es más probable que alguien del grupo B se le clasifique erróneamente como riesgo bajo. Estos dos valores nos dan indicios de impacto diferencial negativo sobre el grupo B. En caso de que el grupo B fuese el más desprotegido, entonces dichos resultados deberían ser analizados en su contexto social y operativo para determinar si existe algún tipo de sesgo o discriminación.

De nuevo, es importante recordar que estos parámetros de interpretación se tendrán que acordar con el cliente, durante la fase de elaboración del Plan de análisis. **La interpretación de los resultados de estas medidas, dependerá siempre del caso concreto**. Por ejemplo: no es lo mismo que un grupo vulnerable tenga un 30% más de FNR, que si lo tiene un grupo privilegiado. En el primer caso, el modelo estaría generando una desventaja que podría ser discriminatoria hacia el grupo protegido. En el segundo caso, puede considerarse una forma de discriminación positiva e, incluso, necesaria.

Por otro lado, pueden darse casos donde una diferencia elevada en los ratios de falsos positivos y negativos, que resulta discriminatoria para un grupo protegido, pueda ser justificada por el funcionamiento de un sistema específico. Por ejemplo, esto sucede cuando el valor de corte para la asignación de riesgo de grupos vulnerables que tiene mucha presencia muestral en relación con un fenómeno, se establece como muy elevada, de forma intencionada, en el diseño del modelo. Esto podría

hacerse para reducir la presencia de este grupo en la asignación de riesgos. Un caso ejemplo podría ser un algoritmo orientado a predecir el riesgo de reincidencia de la población carcelaria en algunos estados de los Estados Unidos, donde la población de riesgo es predominantemente afroamericana. No obstante, cabe realizar un análisis ético y de deseabilidad de aquellos sistemas diseñados en estos términos.

También se podrá evaluar cómo **responde el sistema a datos** de entrada nuevos/diferentes (puede decidirse intercambiar unos datos por otros) **y a órdenes** impuestas por el equipo auditor. Por ejemplo, es posible valorar la precisión del resultado estimado de un algoritmo para un individuo o un grupo, basándose en el análisis de los perfiles de las personas que componen las bases de datos de entrenamiento, y en relación a cómo este se ha comportado con otros individuos o grupos de características comparables. Esto permitiría indicar al cliente hasta qué punto se puede confiar en el algoritmo cuándo este se aplique a un caso concreto.

### *3.3.4.2 LA AUDITORÍA CUALITATIVA*

De forma paralela, se desarrollará la **parte cualitativa** del análisis, **igualmente necesaria para la validación** del algoritmo. Dado que, como ya se ha indicado, la auditoría es un proceso cíclico, esta parte de análisis cualitativo se va retroalimentando y aporta información esencial para el análisis cuantitativo. Esta, consiste en recopilar, analizar e integrar en el análisis y la interpretación de los resultados, toda aquella información que sirva para valorarlo de manera holística. Esta información podrá recogerse y analizarse mediante la **revisión de literatura académica** al respecto, u otros documentos de interés. También a través del **intercambio de información** con las distintas partes implicadas en el diseño, desarrollo e implementación del algoritmo y las partes afectadas directa e indirectamente por el mismo, o mediante la realización e interpretación de los resultados de **entrevistas**, entrevistas en

profundidad, **encuestas**, **grupos de discusión**, **observación** participante o no participante, **estudios etnográficos**, **paneles de expertos**, etc.

El análisis cualitativo de un sistema algorítmico se centra fundamentalmente en examinar que **los principios de cumplimiento ético y legal, aceptabilidad, deseabilidad y protección de los datos personales, se cumplen** en el contexto concreto del sistema. Para ello, estudia los **objetivos y los usos** del algoritmo, la **protección o desprotección** de los individuos y los grupos afectados por el mismo, así como el **cumplimiento de las normas políticas, sociales, jurídicas y éticas** aplicables, y su **integración en dinámicas** más amplias. Como se ha explicado antes, dependiendo del tipo de sistema, esto supone (re-)analizar la composición sociodemográfica del grupo objetivo del algoritmo en su marco social de funcionamiento, examinar (o re-examinar) la literatura teórica sobre aquel fenómeno o variable que se desea medir y estudiar la composición de la muestra utilizada para el entrenamiento del algoritmo. Por ejemplo, en el caso de un algoritmo diseñado para predecir el riesgo de quedar en situación de calle en una ciudad concreta, será necesario recoger y analizar datos reales y literatura teórica al respecto de aquellas variables que mejor reflejan la posibilidad de que las personas vivan en la calle en esta ciudad, la cantidad de gente que está en situación de calle y sus grupos de pertenencia relevante, entre otras.

Cabe destacar que, un aspecto imprescindible de la evaluación cualitativa de un algoritmo dado es comprender **cómo y a quién afecta su creación y su uso**. Por ello, es especialmente recomendable recabar información a través de personas, grupos u organizaciones afectadas por el mismo, y conocer sus niveles de satisfacción y posturas con respecto a la utilización de esta técnica en relación con un determinado problema. Este estudio permitirá al equipo auditor plantear mejoras del sistema, basadas en una comprensión más completa de su impacto social.

Asimismo, el análisis cualitativo de un sistema algorítmico incluye examinar **cuál será el papel que este jugará en los procesos** en los que se integra, y también **analizar el perfil, la formación o la satisfacción del equipo** que interactúa con él. Esto implica resolver algunas cuestiones relevantes, que incluyen: ¿Cuáles son los procesos en los que se integra el algoritmo? ¿Tiene el algoritmo un rol adecuado en estos procesos? ¿Se han modificado las rutinas y dinámicas de la organización/organizaciones que utilizan el algoritmo, o se mantienen con respecto a la situación anterior al algoritmo? ¿Cuáles son estas rutinas y dinámicas actualmente? ¿Cuáles son los equipos y los perfiles profesionales que interactúan con el algoritmo? ¿Están adecuadamente formados y entrenados para utilizar el algoritmo de manera proporcionada? Para ello, se recopilará información sobre los **roles y los perfiles profesionales** de los miembros del equipo, sus **responsabilidades** en relación al funcionamiento del sistema, el **entrenamiento** proporcionado a este equipo, al igual que otros aspectos, incluyendo: si su nivel de **confianza** en el sistema es adecuado[21], si todos los trabajadores que interactúan con el algoritmo lo utilizan de una manera **unificada**, o si por el contrario aplican sus resultados de manera dispar, o datos de **satisfacción** interna/externa con el sistema. Asimismo, cabe preguntarse si las **condiciones de privacidad**, los **principios de protección de datos** y las **medidas de seguridad** se están cumpliendo de manera adecuada y de acuerdo con la normativa jurídica y los códigos deontológicos vigentes.

[21] De acuerdo con el artículo 22 del RGPD, la auditoría debe reflejar si, en el caso de un algoritmo integrado en un proceso de toma de decisiones, permite la intervención humana y sigue primando el criterio profesional de los trabajadores especializados en el asunto, sobre el resultado aportado por el algoritmo. También se recomienda analizar cuál es el peso del resultado del algoritmo en la decisión final.

A continuación se muestra un ejemplo de diagrama que refleja los pasos básicos de utilización de un algoritmo ficticio de adjudicación de recursos:

*Ej.: Diagrama del proceso del uso de un algoritmo de adjudicación de recursos*[22]*:*

| | |
|---|---|
| | Una persona X con unas necesidades específicas, contacta con una institución Y para solicitar un determinado recurso. |
| | Un/a primer/a trabajador/a de Y atiende la solicitud de X, introduce en una base de datos información el/ella y su problemática, acompañada de sus datos personales y atributos sensibles que la definen. |
| | Un/a segundo/a trabajador/a de Y clasifica la solicitud de X, introduce la información en la base del algoritmo y la reporta también a otras dos oficinas especializadas en esta cuestión. |
| | El algoritmo evalúa el conjunto de solicitudes conforme a los recursos disponibles, los datos de la persona, etc. y genera una alerta cuando se asigna un porcentaje de riesgo mayor al 95%. Si la persona no está entre el 5% e mayor riesgo, el algoritmo no emite una alerta. |
| | Un/a tercer/a trabajador/a evalúa las propuestas realizadas por las dos oficinas especializadas junto con el resultado e valoración el algoritmo, en el caso de que este haya emitido una alerta, y toma una decisión final. En el caso de suponer la atribución de un recurso, es comunicada al resto del equipo. |

[22] En este caso de ejemplo, vemos que la persona solicitante va a ser objeto de una decisión parcialmente automatizada por parte del algoritmo de asignación de recursos, por lo que, entre otras cosas, tiene derecho a conocer el proceso y los motivos por los que se toma la decisión de asignarle o no un recurso determinado y, en su caso, recurrir para que su caso sea evaluado íntegramente por un humano.

La auditoría también tiene el propósito de combatir la **opacidad algorítmica**, sugiriendo medidas de **transparencia**, que ayuden a explicar el funcionamiento, las debilidades y fortalezas y los resultados del algoritmo. Esto incluye hacer público, por ejemplo, cuáles son las variables, intersecciones o variables proxy, más determinantes para el sistema, o cuya variación habría afectado más a sus resultados. También se deben comunicar de manera adecuada las **responsabilidades y las medidas de rendición de cuentas** ligadas a los resultados del diseño, desarrollo e implementación del sistema[23].

Los resultados de los análisis cualitativos y cuantitativos serán plasmados e interpretados en el Informe de auditoría, conforme a los parámetros definidos en el Plan de análisis.

[23] En el apartado de recomendaciones se presentarán prácticas y medidas aconsejadas para la mitigación de sesgos, la remodelación o el perfeccionamiento del sistema, y también para la mejora de sus aspectos cualitativos, como el correcto uso del sistema, o la implementación de medidas de responsabilidad y transparencia.

### 3.3.5 INFORME DE AUDITORÍA: EXPLICACIÓN, INTERPRETACIÓN DE RESULTADOS, RECOMENDACIONES Y CONCLUSIONES DE LA AUDITORÍA.

Tras la ejecución del análisis se elaborará un **Informe de auditoría**, que deje constancia del proceso realizado, así como del **cumplimiento legal**

**y ético, la precisión, la aceptabilidad y la deseabilidad** del modelo en base la interpretación de los resultados. Como se ha indicado en secciones anteriores, los parámetros de interpretación de resultados serán acordados con el cliente durante la fase de elaboración del Plan de análisis. Esta interpretación se puede hacer a tres niveles:

- con respecto a la base de datos utilizada para su desarrollo, entrenamiento e implementación,

- con respecto los objetivos de la creación e implementación del algoritmo,

- y con respecto al contexto real en el que se inscribe.

Asimismo, se valorará la **adecuación de los resultados al Plan de análisis** inicial y la interpretación de los resultados del análisis **según los objetivos, las hipótesis y las preguntas de investigación establecidas**. También se aportarán **recomendaciones** finales y posibles **medidas de mitigación** de errores para la mejora del desarrollo o la implementación del algoritmo, o para remodelaciones futuras del sistema.

El resultado de la auditoría debe mostrar de forma clara y fácilmente comprensible el **nivel de riesgo** del sistema, preferiblemente en relación a cada una de las variables o grupos observados de manera predominante

en el análisis. En el apartado de Anexos de esta Guía (Anexo 3) se reproduce un ejemplo de tabla de valoración de riesgo.

Esta valoración debe estar claramente documentada en relación a los resultados de las métricas aplicadas en la parte cuantitativa y los análisis cualitativos realizados.

El Informe de auditoría debe tener una extensión adecuada a la complejidad, el tiempo de duración y los contenidos del análisis de la auditoría y debe incluir, al menos, información sobre:

- el título del proyecto y el nombre del sistema auditado;
- la fecha del informe de auditoría[24] y el nombre de los autores del informe/estudio, si corresponde;
- la responsabilidad del equipo auditor en relación con la calidad del Sistema,
- la explicación y contextualización del caso concreto de estudio, incluyendo toda aquella información relevante sobre el algoritmo auditado, recogida como parte del listado de requisitos iniciales, pero también sobre el marco social, económico, organizacional, legal, ética o tecnológica en los que se integra el sistema;

[24] La fecha del informe es aquella en que se han completado los procedimientos de auditoría necesarios para formarse una opinión sobre el nivel de riesgo del sistema.

- la metodología y los pasos del proceso de análisis del algoritmo, incluyendo información sobre los términos y plazos de la auditoría, consensuados con el cliente en el Plan de análisis;

- los resultados del análisis cualitativo y cuantitativo realizado durante la auditoría, organizados y representados de una forma visual y ordenada;

- una explicación razonada y argumentada de la interpretación de los resultados, incluyendo la valoración del sistema (por partes);

- las conclusiones generales y específicas que se extraen de la interpretación de resultados. Incluyendo los aspectos positivos y negativos del algoritmo auditado;

- un listado de prácticas y medidas recomendadas para la mejora del sistema, elaboradas en relación al caso concreto del algoritmo, que sean operativas, claras e implementables;

- un listado de las referencias utilizadas para la elaboración del informe;

- un apartado de anexos (si corresponde).

En el apartado de Anexos de esta Guía (Anexo 2), se incluye, a modo de ejemplo, un modelo de índice de contenidos más desarrollado, para la realización de un Informe de auditoría. Este ejemplo permitirá al lector hacerse una idea más acotada de los contenidos que debe incorporar un informe de auditoría.

# IV. RECOMENDACIONES PARA LA MEJORA DE LOS SISTEMAS TRAS LA REALIZACIÓN DE UNA AUDITORÍA

Cuando se utiliza un algoritmo que trata datos personales o sensibles, o que puede tener un impacto sobre la vida de una persona o un grupo de

personas es recomendable la realización de auditorías algorítmicas.

Una auditoría algorítmica debe señalar los aspectos positivos y negativos del sistema auditado y, especialmente en el caso de aquellos negativos, **aportar recomendaciones que permitan al cliente u organización mejorar el algoritmo o su implementación**. Al igual que el resto del proceso de auditoría, las recomendaciones específicas para la mejora de un algoritmo, **dependerán del caso concreto y de los resultados particulares de los análisis** de precisión, deseabilidad o aceptabilidad del sistema llevados a cabo.

La realización de este tipo de auditorías permite, además de **identificar aquellos posibles incumplimientos normativos que deben ser subsanados**, los aspectos que pueden ser mejorados y optimizados para conseguir que el algoritmo sea más **explicable, más transparente, más predecible y más controlable**. Su práctica es recomendable a aquellos responsables que incorporen a sus tratamientos algoritmos de impacto social, ya sean organismos públicos o entidades privadas, en cuyo caso además contribuirá a fomentar la responsabilidad social corporativa.

En esta sección, se presentan algunos **ejemplos de recomendaciones específicas** que se podrían presentar tras la realización de una auditoría para la mejora de sistemas, con el objetivo de ayudar al lector de esta Guía a comprender en qué consiste esta cuestión[25]. Es importante tener en cuenta que las recomendaciones que se realicen, estarán siempre determinadas por la **fase de desarrollo** en la que se encuentre el sistema. Aquí se recogen diversas recomendaciones que podrían corresponder a diferentes fases, **divididas en las secciones** que se detallan a

[25]De nuevo, estas recomendaciones se basan en la experiencia previa del equipo auditor de Eticas Research and Consulting y la Universitat Pompeu Fabra.

continuación. Estas incluyen consejos concretos orientados a que el tratamiento de los datos realizado por el algoritmo respete las **normas y los principios de protección de datos**. Asimismo, se subraya la **importancia de implementar y reforzar los mecanismos de transparencia** en la supervisión del funcionamiento del algoritmo, que garanticen el cumplimiento de determinadas **obligaciones por parte del responsable** del tratamiento de los datos y que garanticen a los interesados el ejercicio de sus derechos.

## 4.1 RECOMENDACIONES RELATIVAS A LA GESTIÓN DE LOS DATOS Y LA PRECISIÓN DE UN ALGORITMO

### 4.1.1 RESPECTO A LAS BASES TEÓRICAS/METODOLÓGICAS DEL SISTEMA

- En el caso de que se detecten imprecisiones en las **asunciones** de base que fundamentan un algoritmo, se recomendará revisarlas en función de la teoría y los datos disponibles al respecto.

- Asimismo, se recomendará reforzar la **revisión de la literatura académica** sobre aquellos aspectos, variables, contextos que se ven afectados por el sistema, si estos se consideran insuficientes o inadecuados.

- La misma recomendación se aplica para las **bases metodológicas** de creación del algoritmo, en el caso de que estas no se consideren apropiadas, como por ejemplo la forma de recoger los datos de entrenamiento del sistema.

### 4.1.2 RESPECTO A LA BASE DE DATOS

- Revisar la veracidad, fiabilidad y actualización de la **fuente** de procedencia de los datos.

- Examinar la **representatividad** de la muestra de una variable, intersección o un grupo de variables que definen un grupo de análisis, con respecto a unos parámetros determinados o con respecto a la realidad.

- **Minimizar** la recogida de datos, en general, y en particular de aquellos que no son necesarios para el fin del algoritmo o cuya recogida puede estigmatizar a personas o grupos concretos.

- En el caso de que se produzcan **desequilibrios** entre la cantidad de datos que el sistema recoge sobre una determinada variable con respecto a otra, pudiendo dar lugar a desviaciones del sistema. La recomendación de minimización o ampliación de la cantidad de los datos debe incorporar un análisis preciso de compensación, estableciendo la relación entre la cantidad y tipología de datos a ser recogidos/descartados y aquellos necesarios para garantizar la efectividad y eficiencia del sistema en cuestión.

- Si no se han recogido en la base de datos de entrenamiento o testeo del sistema categorías de datos o de **variables que son necesarias para el correcto modelamiento** del algoritmo, se podrá recomendar que se incluyan. En algunos casos, no recoger determinadas variables durante el proceso de entrenamiento del sistema, puede implicar que el algoritmo no las identifique o "aprenda" y no las pueda utilizar en el futuro.

- El caso referido en el punto anterior puede darse en sistemas que necesitan recoger información sobre atributos sensibles para **desempeñar su función**, o para **valorar que el sistema sea preciso con respecto a dichos atributos** (por ejemplo: para controlar que un algoritmo no discrimina por razones de género, debe haber información sobre el género de las personas que componen la base de datos).

- Modificar el **formato de los datos de entrada**, si este no es adecuado porque no representa la realidad que refleja, o en relación con la forma de funcionamiento del sistema (por ejemplo: si valoramos el caso de un algoritmo de procesamiento del lenguaje natural y la forma de funcionamiento de dicho algoritmo no tiene la capacidad de adaptarse a

cambios en las palabras que componen los textos de entrada, es probable que el algoritmo no se comporte de la manera deseada, si los textos de entrada no son esquematizables. En este caso se podrá optar por una forma de entrada de los datos que sea más ordenada, o bien adaptar el funcionamiento del sistema al formato de los datos de entrada).

- Cambiar la **forma de recoger los datos**. Es posible que, en algún caso la manera en la que el algoritmo recoge datos no sea adecuada, y sea recomendable aplicar filtros en la recogida, ampliar o restringir los datos que se recogen.

- **Limpiar o reestructurar la base de datos** y la clasificación de las variables en tipos claramente distinguibles e identificables.

- **Limpiar o reestructurar el diccionario** de la base de datos, si este no es legible o no resulta operativo para comprender la base de datos.

- Revisar si la **distribución o la frecuencia de variables** recogidas en la base de datos son inadecuadas, dado que esto puede provocar desequilibrios del sistema.

### 4.1.3 RESPECTO AL TRATAMIENTO DE DATOS Y VARIABLES

- En el caso de que la base de datos contenga información sobre atributos identificativos de grupos vulnerables, se podrá recomendar que esta información **no se tenga en cuenta para el modelamiento del sistema** (sino únicamente para valorar su precisión), o bien que se monitorice su comportamiento con respecto a estas variables a lo largo del tiempo.

- Estudiar los casos de **variables con muy baja prevalencia en la muestra, que se considera que no pueden ser modelables de forma robusta** y generar alertas cuando el sistema las detecte. Esto se refiere al caso de aquellas variables de las que no se recogen casos o se recogen pocos en la base de datos. También a aquellas variables para las cuales, si se produce una pequeña variación al alza o a la baja de su presencia en la base de datos, darían lugar a un modelo diferente

- En el caso de que un algoritmo **no sea preciso o discrimine a grupos sociales específicos** por su vinculación con un atributo determinado, se recomendará revisar dicho comportamiento o remodelar el sistema para su corrección. En este sentido, es posible recomendar integrar al modelo una o varias variables no contempladas inicialmente, pero detectadas como discriminatorias durante el análisis mediante *proxies* u otros métodos. Esto tiene el fin de tener mayor control sobre dicha variable y que el algoritmo la identifique correctamente como un valor de medición.

- Se recomienda también realizar un análisis de efectividad de estas cuestiones mediante la **comparación de subgrupos afectados** por la cuestión.

- Esto puede ocurrir también con respecto a determinadas **intersecciones** entre variables, supuesto en el que también se recomendará la revisión y posible reconfiguración del comportamiento del sistema respecto a ellas.

- Se recomienda prestar especial atención a la **cantidad de información** recogida en la base de datos sobre aquellos atributos/variables que el sistema subestima/sobreestima, o a las **reglas** por las cuales puede producirse esta situación (en ciertos casos puede ser de forma intencionada).

### 4.1.4 RESPECTO AL FUNCIONAMIENTO DEL ALGORITMO

- Una cuestión relevante puede ser revisar el **nivel de estatismo o de variabilidad** del modelo, en relación con el tipo de datos de entrada que este maneja, la estructura de recogida de datos, el entorno con el que interactúa, la forma de aprendizaje del sistema, etc. Esto incluye evaluar si el sistema puede y debe o no adaptarse a nuevos datos o nuevos tipos de datos de entrada, si puede extraer conclusiones válidas a partir del formato de información que maneja, si puede aprender nuevas relaciones entre datos de entrada y de salida, etc.

- En el caso de que la evaluación indicada en el punto anterior sea negativa, puede recomendarse **cambiar la forma de aprendizaje del sistema**. Es decir, pasar de uno más supervisado, a otro menos supervisado, o a la inversa.

- Una cuestión importante, aunque complicada de prever, es el **comportamiento futuro de un algoritmo**. Esto dependerá en gran medida de los datos con los que el algoritmo interactúe, el ecosistema de retroalimentación que generen estos datos y otros factores relativos al contexto, que sean cambiantes. En este caso, se recomienda **monitorizar en el tiempo el comportamiento del sistema con respecto a factores cuya variación pueda afectar a su funcionamiento en el futuro**. Esta variación puede relacionarse, por ejemplo, con aquellas variables cuya representación en las bases de datos de entrenamiento del sistema es demasiado pequeña o demasiado grande, sin que esto se corresponda con la realidad social, o sabiendo que esta realidad social puede cambiar.

- También se recomienda monitorizar en el tiempo cómo pueden afectar al desarrollo y el funcionamiento del sistema los cambios en el **contexto social, económico, organizacional, medioambiental, etc**. Estos cambios pueden tener efectos sobre las variables objetivo del modelo algoritmo alterando su eficiencia. Por ejemplo, en el caso de que

la población en situación de calle del género femenino aumente abruptamente en una determinada población que es evaluada mediante un algoritmo de asignación de riesgo. Si el algoritmo no es capaz de captar o aprender esta transformación social debidamente podría infravalorar el riesgo de las mujeres y limitar los recursos públicos asignados a las mismas.

- Por último, se recomienda la **realización de auditorías periódicas**, que no se restrinjan a un momento determinado del desarrollo e implementación del algoritmo, sino que permitan evaluar su evolución en el tiempo. La evaluación anual de un sistema suele ser suficiente.

## 4.2 RECOMENDACIONES RELATIVAS AL CUMPLIMIENTO ÉTICO Y LEGAL

- De forma general, el **cumplimiento de los derechos fundamentales a la privacidad y la protección de datos personales** debe ser respetado y, en la medida de lo posible, promovido, tanto en los procesos de diseño, desarrollo e implementación de un algoritmo, como durante el proceso de auditoría. Esto debe ser así también para todos aquellos derechos que se puedan ver afectados en el caso concreto de un algoritmo.

- El desarrollo e implementación de un algoritmo, y toda auditoría algorítmica que se realice deben prestar especial atención a aquellos **aspectos de un algoritmo que pueda incumplir lo establecido por el RGPD, la LOPDGDD** u otras normas sectoriales o estándares nacionales o internacionales.

- Especialmente, se debe impulsar el respeto a los **principios de tratamiento de los datos**, recogidos tanto en el RGPD como en la LOPDGDD.

- Se recomienda que el desarrollo e implementación de los algoritmos se realice de acuerdo con lo establecido por los **códigos éticos y deontológicos** del sector en el que se implementa.

- También se recomienda aplicar las nociones recogidas en las **guías y catálogos de medidas y buenas prácticas**, emitidas por las autoridades competentes.

- Como parte del desarrollo e implementación de un algoritmo se deben establecer **medidas que faciliten el ejercicio de derechos** de la ciudadanía.

- Todo método de tratamiento de datos, ya sea convencional o mediante el uso de un algoritmo, debe quedar reflejado como actividad en un **Registro de Actividades del Tratamiento**. Se recomienda así que las auditorías algorítmicas verifiquen que el tratamiento realizado por el algoritmo esté adecuadamente recogido en el RAT y recoja toda la información contemplada en el Artículo 30 del RGPD.

- La recopilación y el tratamiento de datos personales y datos sensibles debe ser especialmente cuidadoso con estas cuestiones, en particular, cuando se refiere a **grupos vulnerables**.

## 4.3 RECOMENDACIONES PARA UNA MAYOR ACEPTABILIDAD Y DESEABILIDAD DEL SISTEMA

### 4.3.1 RESPECTO AL USO DEL SISTEMA

- Durante el desarrollo del algoritmo, y de forma previa a su implementación, se recomienda que se realice un **repaso y un estudio de casos** con los equipos que utilizarán el algoritmo, o que aplicarán sus resultados, de modo que puedan **aportar su opinión y sugerir cambios o mejoras en función de su experiencia**.

- Se recomienda garantizar la **formación y la preparación** necesaria de los trabajadores y las trabajadoras que interactúen con el modelo (directa e indirectamente). Esto ayudará a que el nivel de confianza de los equipos humanos sea adecuado, es decir, ni demasiado, ni muy poco. De este modo, será más fácil garantizar un **equilibrio apropiado entre el criterio profesional y los resultados de un algoritmo**.

- También se aconseja realizar **formación continua** que permita a los y las trabajadores sustituir sus prácticas y protocolos de actuación anteriores por las nuevas dinámicas de interacción con el algoritmo, e interiorizar aspectos importantes de su funcionamiento técnico del sistema, su alcance y limitaciones.

- Además de las actividades formativas, se recomienda **recoger datos de satisfacción**, tanto de los **trabajadores y las trabajadoras** que interactúan directamente con el algoritmo, como de aquellas **personas a las que afectan sus resultados**, si esto es posible.

- Es especialmente recomendable recoger también datos de la **satisfacción y afectación de las personas interesadas**, sobre las que tiene un impacto el desarrollo y el uso del algoritmo, especialmente si

estas son individuos o grupos vulnerables y contar con su colaboración durante el proceso de audición del sistema, especialmente a la hora de recomendar mejoras.

- En el caso de que el algoritmo afecte a **individuos o a grupos vulnerables**, también es recomendable recoger datos de satisfacción de **organizaciones sociales** u otro tipo de instituciones que trabajen con estas personas y, al igual que en el caso anterior, contar con su colaboración durante el proceso de audición del sistema, especialmente a la hora de recomendar mejoras.

- En aquellos casos en los que la situación lo permita, será de especial relevancia **comparar estos datos de satisfacción con respecto al sistema utilizado de forma previa al algoritmo** para dar respuesta al mismo problema o a un problema similar. Esto permitiría valorar aspectos cruciales de la deseabilidad y la aceptación del sistema empleado.

- También se recomienda evaluar los **puntos de debilidad y fortaleza, amenazas y oportunidades** del modelo algorítmico, con respecto al proceso de resolución del problema anterior al empleo de este.

- Una vez que un algoritmo se pone en marcha, es importante conocer **cómo y en qué casos lo utilizan las personas y los equipos humanos que interactúan con él**.

- Una cuestión clave a este respecto consiste en saber si sus **resultados se aplican de una manera unificada**, o si su interpretación o utilización difiere en función de la persona que interactúe con él.

- En relación a esto, se recomienda **recopilar información sobre cuál es el baremo humano** utilizado para interpretar y utilizar los

resultados provistos por el algoritmo utilizado. La definición de este baremo debe ser clara, dado que este determina el peso que pueda tener la validación humana del resultado de un algoritmo en la toma de una decisión u otro proceso relevante.

- En el caso de los algoritmos que se utilizan para ayudar en la toma de decisiones, como es el caso de los algoritmos de clasificación, de precisión o de recomendación, también se recomienda **recoger datos sobre el resultado del algoritmo y la decisión final tomada por la persona que interactúa con él**, de cara a tenerlo en cuenta para posibles ajustes del modelo en el futuro. Esta recogida de información deberá tenerse en cuenta desde el diseño de las dinámicas en las que se integra el algoritmo y debe realizarse también de acuerdo con los principios de tratamiento de los datos especificados en esta Guía.

- También se recomienda establecer un proceso por el cual se especifique qué se debe hacer y cómo se debe informar a los responsables del desarrollo e implementación del algoritmo, **en el caso de que la validación humana indique que se debe proceder de una manera contraria o significativamente diferente a la indicada por el algoritmo**.

- Con respecto a las **dinámicas y procesos** en los que se integra el sistema, se recomienda explicitar en qué han cambiado con respecto a antes de la existencia del sistema y cómo se han adaptado a ello las instituciones que lo implementan.

- Lo mismo ocurre para los **objetivos específicos** perseguidos con su uso.

### 4.3.2 RESPECTO A LAS MEDIDAS DE TRANSPARENCIA Y LOS MECANISMOS DE RESPONSABILIDAD Y RENDICIÓN DE CUENTAS

- Se recomienda **explicitar** de cara a las personas que interactúen con el modelo o a las que este pueda afectar, así como a la ciudadanía en general, información clara sobre, al menos, los **objetivos del algoritmo, sus funciones, el tipo de datos que trata, cómo los utiliza, cómo se utilizan los resultados del algoritmo, o con quién se comparten estos datos**.

- Hay que considerar que los usuarios no se encuentran, en muchos casos, en condiciones de asimilar o comprender esta información, por lo que debe exponerse de **forma concisa, sencilla y, a ser posible, visual**.

- En los casos en los que se utilicen **variables proxy**, se recomienda describir de la forma más precisa posible qué variables proxy está capturando el sistema, cómo las combina y para qué lo hace.

- Respecto a los **datos de precisión del modelo** se debe indicar cuáles son los parámetros y valores de corte para que el modelo tome en cuenta determinadas variables a la hora de aportar resultados que sean significativos. Explicar esto de forma clara a los afectados. Esto es aplicable también en el caso de la realización de las auditorías, en cuyo caso la transparencia del proceso debe ser la mayor posible, tanto hacia el cliente, como hacia las partes interesadas y, si corresponde, el público general (en el caso de que la auditoría se haga pública por acuerdo de todas las partes).

- En el desarrollo y la implementación de un algoritmo debe recogerse de forma explícita cuál es la **distribución de responsabilidades** al respecto,

de forma que sea claro quién decide qué y quién asume la responsabilidad de los resultados del desarrollo, implementación y uso de un algoritmo, especialmente cuando estos puedan ser negativos.

En ciertos casos, la precisión del sistema viene predeterminada por **medidas de evaluación** realizadas por las personas u organizaciones que desarrollan o implementan el sistema. Es recomendable que estas sean **claramente explicadas** al público. En el caso de que se detecte que esta es una medida inadecuada, se aportarán otras **medidas complementarias** y se recomendará **cambiar o complementar** esta forma de medir la precisión del sistema.

# V. ANEXO

# 5.1 Anexo 1: Glosario

A los efectos de la presente guía, resulta útil definir previamente una serie de conceptos relevantes para la comprensión de la metodología de auditoría algorítmica.

### 5.1.1 ALGORITMO

Como se ha avanzado en el apartado de Introducción, en esta guía, se utiliza la palabra "algoritmo" desde su concepción en el ámbito de la ciencia de la computación. Desde esta perspectiva, un algoritmo consiste, en un conjunto de instrucciones o reglas definidas y no-ambiguas, ordenadas y finitas que permite, típicamente, contestar una pregunta, tomar una decisión, solucionar un problema, realizar un cómputo, procesar datos o llevar a cabo alguna tarea. Estos procedimientos computacionales toman uno o varios valores de entrada y generan uno o varios valores de salida, por lo tanto son instrumentos que no intentan establecer un vínculo causal entre una variable específica y su efecto, sino que producen un resultado.

A menudo, los algoritmos se implementan en los procesos de toma de decisiones[26], para la clasificación de ítems, o para la predicción de sucesos. En el presente, la palabra "algoritmo" suele utilizarse en referencia a procesos computacionales automatizados, llamados Algoritmos de Aprendizaje Automático, que son los más implementados durante las dos últimas décadas. En este glosario se explican las características principales de los algoritmos de aprendizaje automático según su forma de aprendizaje.

---

[26] Es importante subrayar que, el uso de algoritmos en los procesos de toma de decisiones, debe cumplir una función complementaria, y no sustitutoria de la decisión humana, especialmente en aquellas decisiones que puedan afectar significativamente la vida de las personas, en cumplimiento del Artículo 22 del RGPD.

### 5.1.2 ALGORITMO CON IMPACTO SOCIAL

En general, en esta Guía se considera que el uso o la implementación de un algoritmo es especialmente proclive a tener un impacto social, cuando este **maneja datos personales** (o datos cuya identidad vinculada es deducible), toma **decisiones o contribuye a tomar decisiones que pueden tener efectos significativos** sobre el funcionamiento social o la vida de las personas. Estos efectos pueden ser **positivos o negativos**. No obstante, de manera general, cuando se habla de impacto social se refiere a aquellas **afectaciones consideradas negativas**. En el caso de los algoritmos, estos efectos negativos suelen **vincularse a formas de sesgo o discriminación**. En este sentido, los algoritmos **pueden reproducir o reforzar desigualdades existentes o generar otras nuevas**, perjudicando así a personas o grupos vulnerables.

Del mismo modo, es importante tener en cuenta que un algoritmo diseñado para un servicio o producto concreto, bajo medidas razonables y prudentes para cumplir una función determinada, **puede tener un efecto perjudicial desde un punto de vista ético, social e incluso legal**. Esto tiene que ver con los altos niveles de imprevisibilidad de estos sistemas.

### 5.1.3 ALGORITMO DE APRENDIZAJE SUPERVISADO *(supervised learning)*

Aquel algoritmo en donde los humanos actúan como "instructores" del mismo. Es decir, introducen datos de entrenamiento en el sistema, que contienen los datos de entrada y también los datos de salida "correctos" para esos datos de entrada. Estos datos de salida "correctos" son datos etiquetados. El algoritmo debe reproducir este "patrón" en futuras ocasiones, para producir nuevos datos de salida, siguiendo la misma lógica. El objetivo de este tipo de algoritmos es, precisamente, "modelar" el "comportamiento" del sistema.

### 5.1.4 ALGORITMO DE APRENDIZAJE NO SUPERVISADO *(unsupervised learning)*

Aquel algoritmo en donde los humanos no actúan como "instructores" del mismo, ya que: el algoritmo trabaja con datos no etiquetados. Los humanos no entrenan al algoritmo, como en el caso del aprendizaje supervisado. Este tipo de algoritmos se diseñan para ser capaces de detectar patrones y reglas latentes en los datos y para resumir y agrupar las unidades de información que conforman los datos. Por lo tanto, son especialmente útiles en aquellos casos en los que le persona (desarrollador/a u responsable de una organización) no ha definido qué busca en los datos.

### 5.1.5 ALGORITMO DE APRENDIZAJE SEMI-SUPERVISADO (*semi-supervised learning*)

Aquel algoritmo que se encuentra a medio camino entre los supervisados y los no supervisados. Contienen algunos datos de entrada etiquetados pero, generalmente, la mayoría no lo están. De este modo, los datos no etiquetados representan una fuente importante de información para el modelado del sistema, pero se complementan con procedimientos automáticos. Estos algoritmos se consideran más adecuados para la construcción de modelos, dado que, se basan en patrones generados e introducidos por personas, al mismo tiempo que los modifican, aumentando el conocimiento humano experto.

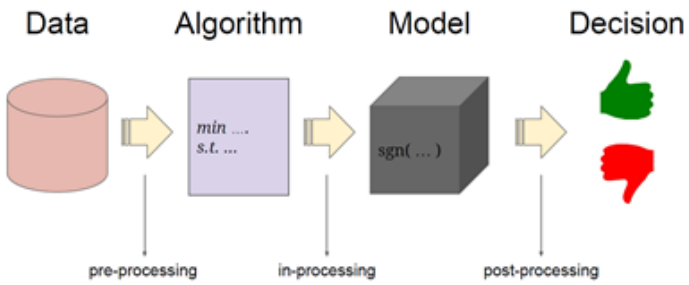
### 5.1.6 ALGORITMO DE APRENDIZAJE POR REFUERZO (*reinforcement learning*)

Algoritmo diseñado para observar la interacción del sistema con su entorno, y aprovecharla para mejorar el funcionamiento del algoritmo. En el proceso de aprendizaje, el sistema analiza y valora diferentes posibles actuaciones, con el objetivo de determinar de forma automática la más idónea dentro de un contexto específico. La señal de refuerzo (*reinforcement signal*) consiste en una retroalimentación simple que el sistema toma como "recompensa" y permite determinar cómo de

"adecuado" es un determinado comportamiento. Esto puede suponer, bien maximizar las virtudes del modelo, bien minimizar sus riesgos, sesgos o efectos indeseables.

### 5.1.7 CICLO DE VIDA DE UN ALGORITMO

El desarrollo e implementación de un algoritmo tiene diferentes fases, representadas en el gráfico a continuación. En primer lugar, se reúne una base de datos que servirán para el entrenamiento y el testeo del sistema. En segundo lugar, se programa el código del algoritmo, que posteriormente se entrena, para generar el modelo algorítmico. Este se prueba previamente a su implementación final.




Fuente: Hajian, Bonchi y Castillo (2016)[27].

En cada fase del desarrollo del algoritmo, las funciones de la auditoría pueden variar. Por eso es importante determinar en qué fase de desarrollo se encuentra el algoritmo y, en función de esto, establecer cuáles son los análisis que se podrán llevar a cabo. La auditoría de un algoritmo se puede concebir desde tres enfoques, en función de estas etapas: en la fase de

[27] El gráfico es una adaptación del texto del artículo de Hajian, Bonchi y Castillo (2016), utilizado a menudo por uno de sus autores, Carlos Castillo.

pueden detectar limitaciones del diseño del algoritmo y proponer medidas que eviten la discriminación; en la fase de **post-procesamiento**, se **pre-procesamiento** se pueden identificar y corregir cuestiones relativas a la base de datos de entrada; en la fase de **procesamiento** se podrán proponer mejoras de modificación de los resultados de los modelos desarrollados (Hajian, Bonchi y Castillo, 2016).

### 5.1.8 DATOS ANÓNIMOS, ANONIMIZACIÓN

Siguiendo las especificaciones proporcionadas por el RGPD (considerando 26), en esta guía se considera que es anónima aquella "información que no guarda relación con una persona física identificada o identificable". Por lo tanto, por anonimización se entiende el proceso encaminado a convertir los datos en anónimos, de manera que una persona no sea identificable a través de ellos.

### 5.1.9 DATOS DE ENTRADA

Los datos de entrada (input *data*) son aquellos que introducen en el algoritmo para ser procesados por el mismo.

### 5.1.10 DATOS DE SALIDA

Los datos de salida (*output data*) son aquellos que resultan del procesamiento algorítmico de los datos de entrada.

### 5.1.11 DATOS ETIQUETADOS

Los datos etiquetados (*labelled data*) son aquellos datos que se introducen en un algoritmo, vinculados a una determinada información de salida. Las etiquetas en los datos permiten al sistema conocer el contenido de estos datos. Un ejemplo de ello sería la identificación de los temas o los atributos contenidos en un fragmento de texto, en el caso de un algoritmo dedicado a esta función. Por ejemplo: dado un determinado texto a un algoritmo de adjudicación de recursos, los datos etiquetados

indicarían que el texto se refiere a una persona de género femenino, con un problema de falta de alimentación.

### 5.1.12 DATOS PERSONALES

En esta Guía se utiliza la definición de "datos personales" proporcionada por el Reglamento General de Protección de datos (Artículo 4. 1). Esto es: "**toda información sobre una persona física identificada o identificable** («el interesado»).

Por "persona física identificable" se entiende "toda persona cuya identidad pueda determinarse, directa o indirectamente, en particular mediante un identificador, como por ejemplo un nombre, un número de identificación, datos de localización, un identificador en línea o uno o varios elementos propios de la identidad física, fisiológica, genética, psíquica, económica, cultural o social de dicha persona;"

### 5.1.13 DATOS SENSIBLES

Al igual que en el caso anterior, la definición de datos o "atributos sensibles" que se maneja en esta guía, viene determinada por aquellos tipos de datos personales a los que, por su naturaleza y por ser especialmente sensibles en relación con los derechos y las libertades fundamentales, el RGPD confiere una especial protección (Artículo 9 y considerandos). Se entiende que, por defecto, datos sensibles son todos aquellos que pertenecen a las denominadas "**categorías especiales de datos personales**" por el RGPD. A saber, aquellos "datos personales que revelen el origen étnico o racial, las opiniones políticas, las convicciones religiosas o filosóficas, o la afiliación sindical, y el tratamiento de datos genéticos, datos biométricos dirigidos a identificar de manera unívoca a una persona física, datos relativos a la salud o datos relativos a la vida sexual o las orientación sexuales de una persona física."

Otros datos que, por su naturaleza, requieren una especial protección, son los datos personales relativos a **condenas e infracciones penales**. El RGPD limita su tratamiento (Art. 10), y establece garantías especiales como la realización de una evaluación de impacto (Art. 35).

### 5.1.14 DISCRIMINACIÓN ALGORÍTMICA

La discriminación algorítmica se refiere al tratamiento desigual proporcionado por un algoritmo a una persona X, con respecto a otra persona Y, debido a un atributo de X, especialmente si ese es un atributo protegido (véase la definición anterior). Esta circunstancia no implica, necesariamente, que la discriminación sea negativa o desventajosa, sino que puede ser también positiva o ventajosa. Esto dependerá de cómo se interpreten los resultados desde el punto de vista ético y social, en un contexto de terminado. Un ejemplo de ello sería una forma de discriminación que afecte positivamente a un grupo protegido o vulnerable (por ejemplo: las personas discapacitadas), al proporcionarle significativamente más recursos que a un grupo privilegiado (por ejemplo: las personas no discapacitadas)[28].

### 5.1.15 DISCRIMINACIÓN ESTADÍSTICA

La discriminación estadística se refiere a la discriminación grupal basada en **un hecho que es estadísticamente relevante**. Esto puede darse, por ejemplo, en el caso de un algoritmo dedicado a la predicción, que utiliza datos sobre probabilidades que proceden del mundo real (y que son estadísticamente relevantes), pero cuyo uso da lugar a un tratamiento desventajoso hacia cierto grupo o colectivo social vulnerable.

---

[28] Las definiciones de discriminación y sesgo presentadas en esta guía se basan, principalmente, en el trabajo realizado por Barocas y Selbst (2016), Baeza-Yates (2018), Castillo (2018), Cowgill (2019), Hajian, S., Bonchi, F., y Castillo, C. (2016), Lippert-Rasmussen (2013), Pedreschi et al. (2008). También en su interpretación para trabajos previos publicados por Eticas Research and Consulting.

Un ejemplo real de ello es el caso de un algoritmo dedicado a la predicción de reincidencia, que se demostró discriminatorio por su uso de la información relativa a los casos de reincidencia entre las personas de piel negra[29].

### 5.1.16 DISCRIMINACIÓN GRUPAL

Esta forma de discriminación algorítmica se refiere a aquella discriminación que afecta a una persona a causa de su pertenencia a un grupo socialmente identificable o protegido. Es decir, un grupo relevante, fundamentalmente en el tejido social y económico.

### 5.1.17 GRUPOS PROTEGIDOS Y/O VULNERABLES[30]

El concepto de grupos protegidos tiene particular importancia para la metodología de auditoría algorítmica de esta Guía parte de una definición de grupos en situación de vulnerabilidad o grupos protegidos fundamentales, que se definen por la pertenencia a ellos de personas que comparten uno o varios de los siguientes atributos protegidos[31]:

- Niños y ancianos (**edad**).
- Tenencia de una **discapacidad o una enfermedad** física o mental.
- Género (**mujer**) o **reasignación** de género.
- Orientación **sexual** (LGTBIQ+).
- Origen étnico o racial, color de piel, ascendencia, condición nacional o inmigrante u otros datos relativos al origen de la persona (condición **racial**).

- Mujeres **embarazadas.**

- **Creencias** u opiniones políticas, religiosas o filosóficas.

- Afiliación **sindical.**

- Información **genética**, **biométrica** o relativa a la **salud.**

- Propiedad o recursos materiales, situación socioeconómica y clase social (condición **socioeconómica**).

- Información sobre **condenas e infracciones penales**.

Esta no es una clasificación exhaustiva, y debe adaptarse o modificarse, en función de cada contexto. Los grupos protegidos se definirán de manera dinámica durante el proceso de auditoría. En el apartado de metodología se retomará esta cuestión.

---

[29] Para más información sobre este ejemplo, relativo al caso del algoritmo COMPAS, consúltese la siguiente página web: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Cabe tener en cuenta, además, que de producirse un caso similar en Europa o que trate datos europeos, de acuerdo con el Artículo 10 del RGPD, el uso de estos datos, relativos a condenas o infracciones penales debería estar adecuadamente informado a las autoridades competentes y tener una base de legitimación.

[30] Cabe destacar que, si bien los términos grupo protegido y grupo vulnerable se utilizan indistintamente a lo largo de este texto y poseen una gran similitud, poseen algunas diferencias. Mientras los grupos vulnerables, ya señalados en el documento y contemplados por diversas normativas, hacen referencia a una serie de colectivos en situación de menor poder o autonomía, la idea de grupo protegido implica la consideración activa y especial de este grupo en el contexto del análisis algorítmico u otro tipo de evaluación social.

[31] Esta clasificación se ha elaborado fundamentalmente según los artículos 6, 9 y 10 del RGPD, los relativos considerando y a la Carta Europea de los Derechos Fundamentales y otros textos de relevancia. Los grupos desfavorecidos pueden definirse en relación con los atributos mencionados en el Artículo 21 (No discriminación), de la Carta Europea de los Derechos Fundamentales: “sexo (y género), raza, color, origen étnico o social, características genéticas, idioma, religión o creencia, opinión política o de cualquier otro tipo, pertenencia a una minoría nacional, propiedad, nacimiento, discapacidad, edad u orientación sexual”.

### 5.1.18 RESPONSABILIDAD Y RENDICIÓN DE CUENTAS

Los algoritmos **no son entes autónomos, sino que carecen de intencionalidad y de voluntad**, **no se les pueden atribuir responsabilidades** con respecto a las normas sociales, éticas o jurídicas.

La **responsabilidad algorítmica** es, por lo tanto, de aquella/s persona/s o grupos de personas u organizaciones que determinan de forma directa los fines y los medios utilizados para el diseño, desarrollo e implementación del algoritmo, que realizan acciones con unas intenciones concretas y unas consecuencias significativas, especialmente cuando estas consecuencias tienen efectos negativos sobre la vida de otro/s. La responsabilidad algorítmica define la relación entre la parte responsable del sistema algorítmico y la parte afectada por el mismo.

La **rendición de cuentas** se refiere a la asunción de esta responsabilidad algorítmica por parte de una persona, grupo u organización. Se refiere a la obligación de reconocer y aceptar las consecuencias del funcionamiento de un algoritmo, así como de reparar y satisfacer a las personas afectadas por el mismo. También se refiere a la responsabilidad de prevenir y evitar posibles consecuencias indeseables en el futuro. La rendición de cuentas, por lo tanto, puede ser **retroactiva** (en relación a acciones pasadas) o **prospectiva** (relacionada con acciones futuras). Establece un vínculo entre los agentes y los pacientes de las consecuencias de un algoritmo y organiza las relaciones sociales, en torno a los procedimientos necesarios para su diseño e implementación.

### 5.1.19 SESGO ALGORÍTMICO

El sesgo algorítmico se produce en aquellos casos en los que un determinado modelo algorítmico basado en datos produce repetidamente resultados no deseados por las personas que desarrollan, crean y entrenan el sistema. Con frecuencia, pero no siempre, esto se debe a que la recopilación y el uso de datos de entrenamiento están sesgados (sesgo pre-algorítmico). En otras ocasiones, se debe a problemas con la

interacción entre un algoritmo y otros procesos, una vez que el algoritmo es aplicado en un contexto concreto (sesgo post-algorítmico).

En aquellos casos en los que estos resultados no deseados dan lugar a una forma de discriminación sistemática, que produce resultados desventajosos que involucran a uno o más de los llamados grupos protegidos o vulnerables, se considera que existe un **sesgo algorítmico discriminatorio** o que se observa discriminación algorítmica.

### 5.1.20 TAXONOMÍA DE IMPACTO SOCIAL

Este impacto social se refiere a los efectos discriminatorios desventajosos o formas de sesgo discriminatorio, producidas por un algoritmo sobre la vida de las personas, especialmente si estos se generan por motivo de su **pertenencia a uno de los grupos vulnerables** antes mencionados. En este sentido, los tipos de impacto social de un algoritmo pueden clasificarse como formas de discriminación, siguiendo la siguiente taxonomía:

- **racial**, de **género**, **sexual**, relativo al nivel **socioeconómico**, relativo a las condiciones **sociodemográficas** (como la edad), **religioso, político** o relacionado con las creencias **filosóficas**,

- relativo a una **discapacidad** o a una **enfermedad** mental o física.

Asimismo, este impacto social puede referirse a los efectos negativos o discriminatorios producidos por un algoritmo, en tanto que este contribuya a:

- transmitir o reforzar una desigualdad social existente (**reproducción de la desigualdad**);

- desinformar, generar desafección o polarización política, obstaculizar el acceso a ideas diferentes u opuestas y mermar así la calidad democrática (**impacto en los procesos democráticos**);

- o vulnerar el cumplimiento de los derechos fundamentales a la privacidad y a la protección de datos de las personas (**impacto en la privacidad**).[32]

### 5.1.21 VARIABLE

El esta Guía se usa el concepto de variable como variable estadística. Una variable estadística es el conjunto de valores que puede tomar cierta característica de la población sobre la que se realiza un estudio (estadístico) y sobre la que es posible su medición.

[32] Esta es una adaptación de la taxonomía desarrollada por el equipo de Eticas Foundation en su Observatory of Algorithms with Social Impact (OASI): https://eticasfoundation.org/algorithms/es/.

## 5.2 Anexo 2: Modelo ejemplo de informe de auditoría algorítmica

A continuación se expone un modelo ejemplo de los contenidos tipo que debería incluir un informe final de auditoría algorítmica:

**1. Portada**

**1.1** Título del proyecto.

**1.2** Nombre del algoritmo auditado.

**1.3** Información de la empresa auditora (como el nombre o el logo).

**2. Subportada**

**2.1** Título del proyecto y nombre del sistema auditado.

**2.2** Fecha de informe.

**2.3** Nombre y organización a la que pertenecen los miembros del equipo auditor.

**3. Índice de figuras y tablas**

**4. Introducción**

**4.1** Alcance de la auditoría y ejes acordados en el Plan de análisis

**4.2** Responsabilidad del equipo auditor

**4.3** Entidad auditada e índice de contenidos del informe

**4.4** Definición del problema algorítmico

**4.4.1** Diseño y desarrollo algorítmico y modelo.

**4.4.2** Cómo se usa el algoritmo: procesos, dinámicas y equipos que interactúan con él.

**5. Objetivos y metodología de la auditoría**

**5.1** Objetivo general de la auditoría.

**5.2** Objetivos específicos de la auditoría.

**5.3** Términos, plazos y ejes de análisis acordados en el Plan de análisis.

**6. Discriminación algorítmica, equidad y principios rectores de la auditoría**

**6.1** Discriminación algorítmica

**6.2** Equidad algorítmica

**6.3** Principios rectores de la auditoría: cumplimiento ético y legal (normas aplicables), aceptabilidad, deseabilidad y la protección y gestión adecuada de los datos personales.

**7. Análisis teórico y estado de la cuestión sobre el tema analizado por el algoritmo**

**7.1** Estado del arte sobre el caso concreto.

**7.2** Estado del arte sobre la problemática concreta.

**8. Hipótesis/preguntas de investigación sobre la precisión del modelo**

**8.1** Hipótesis sobre la validación interna del modelo.

**8.2** Hipótesis sobre discriminación algorítmica.

**8.3** Hipótesis sobre la validación interna para grupos.

**8.4** Hipótesis sobre aceptabilidad y deseabilidad del modelo.

**9. Análisis de la composición del conjunto de datos de entrenamiento y grupos**

**9.1** Composición del conjunto de datos, entrenamiento modelo y precisión general en la asignación de riesgos

**9.2** Grupos protegidos dentro en el conjunto de datos

**9.3** Estructura interseccional de los datos de entrenamiento

**9.4** Tratamiento e impacto diferencial por grupos.

**9.5** Tasa de falsos negativos (FNR) y falsos positivos (FPR) por grupo

**10. Análisis de deseabilidad**

**10.1** Información relevante sobre el contexto social, económico, técnico, organizacional en el que se inscribe el modelo, cómo se ha diseñado y cómo se utiliza, cómo se gestionan los datos integrados en el sistema, cumplimiento con la legalidad y las normas éticas aplicables.

**10.2** Realización de entrevistas, grupos de discusión u otros métodos de obtención de información utilizados.

**11. Resultados: interpretación y valoración**

**11.1** Cuantitativos: precisión global identificada en la asignación de riesgos; sesgo y posible discriminación dentro del sistema.

**11.2** Cualitativos: valoración general del modelo y adecuación a los principios rectores.

**12. Conclusiones**

**13. Recomendaciones y posibles ejes de actuación**

**13.1** Precisión general.

**13.2** Discriminación algorítmica.

**13.3** Futuro remodelado.

**13.4** Cuestiones subsanadas durante el proceso de auditoría.

**13.5** Cuestiones no subsanadas.

## 14. Referencias

## 15. Anexos

**15.1** Acuerdo de confidencialidad.

## 5.3 Anexo 3: Ejemplo de tabla de valoración de riesgo

| Análisis de la precisión y la deseabilidad del modelo en la adjudicación de recursos en función de la variable 'género'<br>[Se analiza si el sistema de asignación de recursos desfavorece a las mujeres] | |
|---|---|
| Factores relevantes | Representación de los grupos en la base de datos de entrenamiento del algoritmo.<br>"La variable 'género' no se recoge de forma explícita en la base de datos de entrenamiento, sino que se infiere a través de otras variables proxies." |
| Mediciones realizadas | Disparidad entre tasas de falsos negativos (FNRs).<br>"La tasa de FNs del grupo de género femenino es mayor que la tasa de FNs del grupo de género masculino en un 51%.<br>Es decir, hay muchos más falsos negativos para la adjudicación de recursos a las mujeres que a los hombres." |
| Observaciones en relación a las hipótesis | Observaciones destacadas.<br>"El sistema infraprotege más frecuentemente al grupo vulnerable (mujeres)."<br>"Las variables proxies utilizadas para determinar el grupo de mujeres y de hombres deben explicitarse mejor."<br>"Se recomienda que la variable 'género' se incluya en el modelamiento del algoritmo para su correcta evaluación." |
| Riesgo | **ALTO** |

| Análisis del tratamiento diferencial del sistema por grupos de edad<br>[Se analiza si el sistema trata de manera significativamente diferente al grupo de edad mayor de 65 años] | |
|---|---|
| Factores relevantes | Correspondencia de la representación en la base de datos de entrenamiento con la realidad.<br>"El grupo de edad de más de 65 años tiene una representación del 20% en el censo nacional. Sin embargo, la base de datos de entrenamiento recoge solo un 6% de casos |
| Mediciones realizadas | Tasas de impacto y tratamiento diferencial entre grupos de edad (DI/DT).<br>"El sistema tiende a asignar un riesgo más bajo a las personas de más de 65 años de edad, que a los grupos más jóvenes. Las diferencias más notables se sitúan en torno al 10%" |
| Observaciones en relación a las hipótesis | Observaciones destacadas.<br>"La representación del grupo en la base de datos es baja (solo 1% mayor del 5% recomendable). Debido a esta baja prevalencia, no se puede afirmar que el grupo pueda ser modelado de forma robusta por el sistema. Esto puede estar desviando la precisión del modelo y explicar la leve disparidad."<br>"La representación en la base de datos también es demasiado baja con respecto a su representación en el censo (20% / 3%). Se recomienda revisarlo para mejorar la precisión del modelo." |
| Riesgo | **MEDIO** |

Se muestra un ejemplo de una tabla de valoración del riesgo de un algoritmo ficticio de asignación de recursos, en relación a dos grupos vulnerables afectados: las personas de origen extranjero (inmigración) y las personas mayores de 65 años (edad). Esta tabla recoge información sobre factores relevantes (como la representación del grupo en la base de datos de entrenamiento), los resultados de mediciones relevantes realizadas como parte del análisis cuantitativo y observaciones derivadas del estudio cualitativo del caso, que ayudan a validar o refutar las hipótesis planteadas, y complementan el análisis cuantitativo. El resultado de la valoración realizada en relación con estas cuestiones se muestra en la última columna, "Riesgo".

## 5.4 Anexo 4: Aspectos relevantes del RGPD y LOPDGDD para la auditoría algorítmica

Este anexo de la Guía destaca **aquellos aspectos más relevantes** de la normativa de protección de datos establecida por el RGPD y la LOPDGDD, que conforman la base jurídica de legitimación para las diferentes etapas de la solución que integra un algoritmo.

Estos textos apuntan posibles cuestiones que tener en cuenta o a las cuales se debe dar respuesta cuando se desarrolla o se utiliza un algoritmo que recopile o trate datos personales y a las que, por lo tanto, toda auditoría algorítmica debe prestar especial atención.

En primer lugar, en cumplimiento de lo expuesto por el RGPD, todo algoritmo, entendido este como una herramienta de tratamiento, debe respetar los **principios de tratamiento de los datos**; aspecto que deberá ser evaluado a la hora de realizar una auditoría algorítmica. Estos principios (Art. 5) se refieren a la licitud, lealtad, transparencia, limitación de la finalidad, minimización, exactitud, limitación del plazo de conservación, integridad y confidencialidad en la recopilación y el tratamiento de los datos y la responsabilidad proactiva del responsable del tratamiento de estos datos.

Además, y de acuerdo con el Artículo 6 del RGPD, **para que un tratamiento de datos sea lícito ha de ampararse en alguna de las siguientes bases legitimadoras**: la **persona interesada** (es decir, una persona física identificada o identificable, Art. 4.1.) ha dado su consentimiento para ello; este tratamiento es necesario para la ejecución de un contrato del que esta persona es parte; es necesario para el cumplimiento de una obligación legal o para proteger intereses vitales; se requiere para cumplir una misión realizada en interés público o en el ejercicio de poderes públicos conferidos al **responsable del tratamiento**

(según el Artículo 4.7, la persona física o jurídica, autoridad pública, servicio u otro organismo que, solo o junto con otros, determine los fines y medios del tratamiento); o es necesario para satisfacer intereses legítimos del responsable del tratamiento.

Los Artículos 7 y 8 amplían la información sobre el **consentimiento del interesado** y las condiciones que debe cumplir para que sea considerado válido. Cabe recordar que el RGPD establece una **categoría especial de datos personales**, particularmente sensibles, cuyo tratamiento queda prohibido excepto los supuestos que establece el Artículo 9. En el caso de que los datos personales se refieran a **condenas e infracciones penales**, el tratamiento de los datos solo se podrá realizar bajo la supervisión de las autoridades públicas competentes (Art. 10).

El RGPD establece una serie de **derechos del interesado**, que se deben cumplir cuando se desarrolla e implementa un sistema que utiliza datos personales. Estos son relativos a: la **transparencia** de la información facilitada al interesado, la adecuada **comunicación** con el mismo y a las diferentes modalidades en las que el interesado puede ejercer sus derechos (Art. 12); la información que deberá facilitarse cuando los datos personales se obtengan de la persona interesada (Art. 13) y cuando estos no se hayan obtenido de la persona interesada (Art. 14); el **acceso** del interesado a los datos personales que le conciernen y a información sobre su tratamiento; la **rectificación**, **supresión**, **limitación** del tratamiento, **portabilidad** y **oposición** al tratamiento de los datos (Arts. 15-22); y el tratamiento automatizado de los datos, utilizado en la toma de decisiones (Art. 22). De acuerdo con este último Artículo 22 del RGPD, toda persona tendrá **derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado de datos**, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar. Este Artículo, por tanto, tiene especial interés en el desarrollo de las auditorías algorítmicas.

El capítulo IV del Reglamento se refiere a las **obligaciones generales del responsable** del tratamiento de los datos (Art. 24), entre las cuales figura la responsabilidad de establecer medidas de **protección de datos desde el diseño y por defecto** (Art. 25).

Además, este capítulo define los **roles que deben ser establecidos**, como son los corresponsales del tratamiento (Art. 26), los representantes de responsables o encargados del tratamiento no establecidos en la Unión Europea (Art. 27) o el encargado del tratamiento (Art. 28). Todos ellos **cooperarán con la autoridad de control** que lo solicite (Art. 31). La autoridad de control es aquella autoridad pública independiente establecida por un Estado miembro de la Unión Europea (Arts. 4.21 y 51), con competencias y potestades (Arts. 57 y 58) en materia de protección de datos. Por su parte, la forma de designación, la posición y las funciones del **delegado de protección de datos**, como figura encargada de asistir al responsable del tratamiento, asesorarle y supervisar el cumplimiento de las exigencias impuestas por la normativa, se recogen en los Artículos 37, 38, y 39 del RGPD.

Este capítulo establece además que, cada organización deberá elaborar un **Registro de Actividades de Tratamiento** (RAT) (Art. 30) y detalla cómo se debe realizar. No obstante, corresponde a cada organización, decidir el nivel de segregación o agregación con el que desea registrar los tratamientos de datos de carácter personal que requiere su actividad[33]

[33] Para más información sobre el RAT, se recomienda acceder a las siguientes páginas web de la AEPD: https://www.aepd.es/es/derechos-y-deberes/cumple-tus-deberes/medidas-de-cumplimiento/actividades-tratamiento y https://www.aepd.es/es/prensa-y-comunicacion/blog/elaborar-el-registro-de-actividades-de-tratamiento. También se recomienda consultar la herramienta Facilita 2.0 que la AEPD ha puesto a disposición de los responsables del sector privado de tratamientos de datos de escaso riesgo: https://servicios.aepd.es/AEPD/view/form/MDAwMDAwMDAwMDAwMDI2MjQ5NzUxNTg3NjUyNzE0MTU4?updated=true.

Este capítulo también determina que los datos **deben ser tratados de forma segura**, de manera que se evite el tratamiento no autorizado o ilícito de dichos datos, su pérdida, destrucción o alteración accidental (Art. 32). Esto implica que el responsable del tratamiento, en base a un análisis del riesgo, debe: establecer medidas técnicas y organizativas de seudonimización y cifrado de los datos; garantizar la confidencialidad, integridad, disponibilidad y resiliencia de los sistemas y servicios de tratamiento de los datos; restaurar la disponibilidad y el acceso a los datos en caso de incidentes; y establecer procesos de verificación, evaluación y valoración regulares de las medidas técnicas y organizativas que garanticen la seguridad del tratamiento. Asimismo, se deben tener en cuenta aquellos riesgos específicos y generales que presente el tratamiento de los datos, y tomar medidas para garantizar que cualquier persona autorizada que tenga acceso a los datos solo pueda hacerlo bajo las instrucciones del responsable. En el caso de que se produzcan **violaciones de la seguridad de los datos**, estas deben ser notificadas a la autoridad de control en un máximo de 72h (Art. 33), así como al interesado, cuando sea probable que esta violación ponga en alto riesgo sus derechos y libertades (Art. 34).

El Artículo 35 establece las normas relativas a la realización de **evaluaciones de impacto relativas a la protección de datos**. Estas evaluaciones deben ser realizadas por el responsable del tratamiento de los datos personales, en aquellos casos en los el tratamiento de los datos pueda comportar un alto riesgo para los derechos y libertades de las personas, en particular si utiliza nuevas tecnologías, por su naturaleza, alcance, contexto o fines. Esto incluye, por lo tanto, diversas **formas de tratamiento de datos basadas en el uso de algoritmos**, particularmente aquellas que procesan grandes cantidades de datos personales o sensibles.

Esta evaluación del impacto genera una necesidad de establecer diferentes formas de **responsabilidad proactiva**. Esto implica que el

responsable del tratamiento debe tomar activamente el control y decidir qué hace en cada momento, anticipándose a los acontecimientos. Es decir, esta responsabilidad implica una intervención activa, sea esta de carácter **retroactiva**, abarcando diversas formas de rendición de cuentas, **o prospectiva**, es decir, mecanismos y medidas de anticipación del riesgo. Dicha necesidad requiere que el o la responsable del desarrollo y el uso de los algoritmos que utilizan datos personales, analicen qué datos tratan, con qué finalidades lo hacen y qué tipo de tratamientos llevan a cabo con el objetivo de determinar qué medidas son adecuadas para cumplir con lo dispuesto en el RGPD. Este es un Artículo **especialmente relevante**, dado que está directamente relacionado con la realización de auditorías

algorítmicas, en tanto que uno de sus objetivos principales, como se señalaba antes es analizar e identificar los puntos de tensión que pueden suponer un incumplimiento de la normativa de protección de datos, de cara a ayudar a corregirlos y tenerlos en cuenta como requisitos de diseño en el desarrollo de los algoritmos. Cuando esta **evaluación de impacto** revele un nivel de riesgo alto, el responsable realizará una **consulta a la autoridad de control**, antes de tratar los datos (Art. 36).

Por su parte, la **Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales** (LOPDGDD) complementa y particulariza lo dispuesto por el Reglamento en el caso español, reforzando la importancia de dar cumplimiento a los principios de protección de datos y a la atención del ejercicio de derechos por parte del responsable, a la vez que incluye determinadas disposiciones aplicables a tratamientos concretos, algunos de los cuales pueden apoyarse en el desarrollo de soluciones que hagan uso de algoritmos.

En consecuencia, tanto el RGPD como la LOPDGDD vienen a establecer los principios directores que cualquier tipo de tratamiento, incluidos

aquellos basados en soluciones de Inteligencia Artificial y que utilicen algoritmos, debe respetar definiendo un marco de desarrollo de las actuaciones de los responsables basado en la gestión de los riesgos para los derechos y las libertades de los interesados y la rendición de cuentas, o capacidad de demostrar el cumplimiento de las obligaciones impuestas por la normativa.

# VI. REFERENCIAS


Agencia Española de Protección de Datos (2018). Guía práctica de Análisis de riesgos en los tratamientos de datos personales sujetos al RGPD. Disponible en: https://www.aepd.es/sites/default/files/2019-09/guia-analisis-de-riesgos-rgpd.pdf.

Agencia Española de Protección de Datos (2018). Guía práctica para las Evaluaciones de Impacto en la Protección de los datos sujetas al RGPD. Disponible en: https://www.aepd.es/sites/default/files/2019-09/guia-analisis-de-riesgos-rgpd.pdf.

Agencia Española de Protección de Datos (2019). Guía de Privacidad desde el Diseño. Disponible en: https://www.aepd.es/sites/default/files/2019-11/guia-privacidad-desde-diseno.pdf.

Agencia Española de Protección de Datos (2020). Guía de Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Disponible en: https://www.aepd.es/sites/default/files/2020-02/adecuacion-rgpd-ia.pdf.

Ada Lovelace Institute (2020). *Examining the black box. Tools for assessing algorithmic system*. London: Ada Lovelace Institute.

Angwin, J., Larson, J., Mattu, S., y Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Barocas, S. y Hardt, M. (2017). Fairness in Machine Learning. *NIPS*. Disponible en: https://mrtz.org/nips17/

Barocas, S. y Selbst, A. (2016). Big Data's Disparate Impact", *California Law Review*, 671. Disponible en: http://www.californialawreview.org/wp-content/uploads/2016/06/2Barocas-Selbst.pdf.

Barocas, S.; Hardt, M. y Narayanan, A. (2019). Fairness in Machine Learning. Limitations and Opportunities. Disponible en: https://fairmlbook.org/.

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philos. Technol.*, 31 (4), 543-556.

Castillo, C. (2018). "Algorithmic Discrimination. Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavour. *Proceedings of HUMAINT Workshop*. Disponible en: https://arxiv.org/pdf/1806.03192.pdf

Castillo, C. (2019). "Fairness and Transparency in Ranking", *ACM SIGIR Forum, 52 (1), 64- 71. ACM*.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *arXiv*:1610.07524. Disponible en: https://arxiv.org/abs/1610.07524.

Chouldechova, A.; D. Benavides-Prado, O. Fialko,y R. Vaithianathan, (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency 81, 134-148*.

Centre for Information Policy Leadership (CIPL) (2020). *Artificial Intelligence and Data Protection. How the GDPR Regulates AI*. Washington: Centre for Information Policy Leadership.

Comisión Europea (2018). Communication from the Commision to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions. Artificial Intelligence for Europe {SWD(2018)137final}. Disponible en: https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF.

Comisión Europea (2020). *Commission Report on safety and liability implications of AI, the Internet of Things and Robotics*. Disponible en: https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en.

Comisión Europea (2020). *Communication: A European strategy for data*. Available at: https://ec.europa.eu/info/publications/communication-european-strategy-data_en.

Comisión Europea (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust*. Available at: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Comisión Europea (2020). Communication from the Commision to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European strategy for data. Disponible en: https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf.

Comisión Europea (2020). Commission Report on safety and liability implications of AI, the Internet of Things and Robotics. Disponible en: https://ec.europa.eu/info/publications/commission-report-safety-and-liability-implications-ai-internet-things-and-robotics-0_en.

Comisión Europea (2020). *White Paper on Artificial Intelligence: a European approach to excellence and trust*. Disponible en: https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

Comisión Europea (2020). *Ethics Guidelines for Trustworthy AI*. Disponible en: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top.

Comisión Europea-Grupo de expertos de alto nivel sobre inteligencia artificial. (2018). Ethics Guidelines for Trustworthy AI. Burselas: Comisión Europea. Disponible en: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419.

Diakopoulos, N. (2015). Algorithmic Accountability. *Digital Journalism*, 3(3): 400-403.

Diakopoulos, N., and Friedler, S. (2016). How to Hold Algorithms Accountable, MIT Technology Review, November 2016. Retrieved from: https://www.technologyreview.com/s/602933/how-to-holdalgorithmsaccountable/?utm_content=buffer19bc5andutm_medium=socialandutm_source=twitter.c%E2%80%A6.

Dwork, C. y Ilvento, C. (2018). Individual fairness under composition, *FATML*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., yamp; Zemel, R. (2012). Fairness through awareness, *Proceedings of the 3rd innovations in theoretical computer science conference, 214-226. ACM*.

Eubanks, V. (2018). *Automating Inequality*. New York: St. Martin's Press.

Fumo, D. (2017). Types of Machine Learning Algorithms You Should Know. *Towards data science*. Disponible en: https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Wortman, J.; Hanna Wallach, V. Daumé III, H.y Crawford, K (2020). Datasheets for Datasets. *arXiv:1803.09010*. Disponible en: https://arxiv.org/abs/1803.09010.

Hajian, S., Bonchi, F. y Castillo, C. (2016). Algorithmic bias: From discrimination on discovery to fairness-aware data mining, *Proceedings of the 22nd ACM SIGKDD international conferencie on knowledge discovery and data mining*: 2125-2126.

Heidari, H., Ferrari, C., Gummadi, K.,y Krause, A. (2018). Fairness behind a veil of ignorance: a welfare analysis for automated decision making. En S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, y R. Garnett (eds.). *Advances in Neural Information Processing Systems* (pp. 1265–1276). Montreal QC: Curran Associates, Inc.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., yamp; Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (p. 600). ACM.*

ICO (2015). *Auditing data protection: a guide to ICO data protection audits*. London: ICO. Disponible en: https://ico.org.uk/media/1533/auditing_data_protection.pdf.

ICO (2019). A Guide to ICO audits. Disponible en: https://ico.org.uk/media/for-organisations/documents/2787/guide-to-data-protection-audits.pdf.

ICO (2020). *Guidance on the AI auditing framework* [Draft guidance for consultation]. London: ICO. Disponible en: https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-and-the-turing-consultation-on-explaining-ai-decisions-guidance/.

ICO (2020). *Explaining decisions made with AI*. London: ICO. Disponible en: https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, y Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 33–44.*

Katell, M. Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft (2020). Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 45–55.*

Kroll , J.; Huey , J.; Barocas, S.; Felten, E.; Reidenberg, J.; Robinson, D. y Yu, H. (2017). Accountable Algorithms, *University of Pennsylvania Law Review* (165). Disponible en: https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3.

Lippert-Rasmussen, K. (2013). *Born Free and Equal? A Philosophical Inquiry Into the Nature of Discrimination*. Oxford: Oxford University Press.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D. y Gebru, T. (2019). Model Cards for Model Reporting, *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*: 220-220.

Narayanan, A. (23 de febrero 2018). Tutorial: 21 definitions of fairness and their politics [Abstract and video] *Conference on Fairness, Accountability, and Transparency*, NYC.

Nissenbaum, H. (2001). How computer systems embody values, Computer. *Computer*, 34 (3): 120–119.

Eticas Foundation. *Observatory of Algorithms with Social Impact (OASI)*. Disponible en: https://eticasfoundation.org/algorithms/.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Sánchez-Monedero, J. y Dencik, L. (2018). *How to (partially) evaluate automated decision systems. Technical Report*. Cardiff University. Disponible en: https://pdfs.semanticscholar.org/2a2d/ecaa5181d18911c3cc3c0e69e3ebdb7649dd.pdf.

Solans, D.; Biggio, B.; Castillo, C. (2020). Poisoning Attacks on Algorithmic Fairness. Disponible en https://arxiv.org/abs/2004.07401

Speicher, T.; Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual &Group Unfairness via Inequality Indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 2239–2248*.

Striphas, T. (2012, Feb 1). What is an Algorithm? *Culture digitally*. Disponible en: http://culturedigitally.org/2012/02/what-is-an-algorithm/.

Vedder, A.; Naudts, L. (2017). Accountability for the Use of Algorithms in a Big Data Environment. *International Review of Law, Computers y Technology*, 31(2): 206 – 224

Wachter, S., Mittelstadt, B. y Russell, C. (2020), Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. *SSRN*. Disponible en: https://ssrn.com/abstract=3547922 or http://dx.doi.org/10.2139/ssrn.3547922.

Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 1–18.*

# Guía de Auditoría Algorítmica

www.eticasconsulting.com – info@eticasconsulting.com - +34 936 005 400