agencia española protección datos

Requisitos para Auditorías de Tratamientos que incluyan IA

v. enero 2021

El presente documento ha sido desarrollado en base al estudio realizado por Eticas Research and Consulting SL bajo el encargo y la supervisión de la Agencia Española de Protección de Datos y las revisiones realizadas por expertos del Artificial Intelligence Hub del Consejo Superior de Investigaciones Científicas (CSIC AI HUB), del Observatorio del impacto social y ético de la inteligencia artificial (OdiseIA), de la Asociación Profesional de Cuerpos Superiores de Sistemas y Tecnologías de la Información de las Administraciones Públicas (ASTIC), Grupo de Innovación Docente en Ciberseguridad (CiberGID)-ETSI Informática - UNED y del Centro para el Desarrollo Tecnológico e Industrial (CDTI).

RESUMEN EJECUTIVO

En el presente documento se ha realizado una primera aproximación a un conjunto de controles que podrían incorporarse a las auditorías de tratamientos de datos personales que hacen uso de componentes basados en inteligencia artificial (IA). Es importante señalar que los controles incluidos están concebidos para realizar un análisis de la adecuación del tratamiento desde una perspectiva de protección de datos. Además, se añaden algunas notas metodológicas que pueden resultar propias y características de este tipo de auditorías.

El documento se encuadra en lo propuesto en la Guía de Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial publicada por la AEPD en lo que respecta al cumplimiento efectivo de los principios de protección de datos personales en tratamientos que incluyan soluciones de inteligencia artificial.

La auditoría de los tratamientos de datos personales es una de las herramientas para la evaluación del cumplimiento normativo. En aquellos tratamientos sometidos al RGPD que incluyan componentes basados en IA será necesario que la auditoria contemple controles específicos derivados de las particularidades de estos. La lista de controles desplegada en este documento pretende ser una referencia para que el auditor pueda determinar, tras un análisis previo, aquellos que sea oportuno y pertinente incluir en la auditoría concreta de un tratamiento.

Si bien es cierto que la auditoría de un tratamiento concreto, en particular uno que incluya componentes de IA, podría verificar otros aspectos de este, como su comportamiento desde un punto de vista ético o su eficiencia desde un punto de vista técnico, el presente documento se centrará exclusivamente en aquellos aspectos relativos a la protección de datos personales.

Por otro lado, la auditoría de un tratamiento que, entre otros, incluya componentes basados en IA, no puede ni debe centrarse exclusivamente en los aspectos técnicos específicos de las tecnologías empleadas, sino que ha de cubrir un alcance mucho más extenso que contemple la naturaleza, ámbito, contexto y fines del tratamiento además de los riesgos para los derechos y libertades que el tratamiento en su conjunto pueda representar.

Este documento está dirigido, principalmente, a responsables que han de auditar tratamientos que incluyan componentes basados en IA, así como a encargados y desarrolladores que quieran ofrecer garantías sobre sus productos y soluciones; a los Delegados de Protección de Datos encargados tanto de supervisar los tratamientos como de asesorar a los responsables y por último, a los equipos de auditores cuando se ocupen de la evaluación de dichos tratamientos.

Palabras clave: Inteligencia Artificial, auditoría, componente IA, algoritmos, aprendizaje automático, Machine Learning, decisiones automatizadas, perfilado, datos masivos, Big Data, RGPD, Protección de Datos Personales, *accountability*, transparencia, rendición de cuentas, cumplimiento legal, sesgo, explicabilidad.

ÍNDICE

I. INTRODUCCIÓN

El RGPD establece en su artículo 24 la obligación de aplicar "*medidas técnicas y organizativas apropiadas a fin de garantizar y poder demostrar que el tratamiento es conforme con el presente Reglamento*". Estas medidas han de ser seleccionadas "*teniendo en cuenta la naturaleza, el ámbito, el contexto y los fines del tratamiento, así como los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas*". Cuando sea necesario, dichas medidas estarán además sujetas a un proceso de constante revisión y actualización.

Una de las herramientas para "*garantizar y poder demostrar*" el cumplimiento del RGPD es la realización de auditorías de los tratamientos. En cumplimiento de sus funciones (artículo 39), el delegado de protección de datos supervisaría esas auditorías. Todo tratamiento ha de ser evaluado, con relación a sus fines, así como en el ámbito y contexto en el que se va a desplegar. A su vez, dicha evaluación ha de tener en cuenta la naturaleza del tratamiento y la necesidad de realizar una evaluación de impacto relativa a la protección de datos (artículo 35) por distintas causas, entre ellas, el tratamiento de categorías especiales de datos a gran escala, el uso de nuevas tecnologías, o la evaluación sistemática y exhaustiva de aspectos personales de personas físicas que se base en un tratamiento automatizado, como la elaboración de perfiles.

Con relación a esto último, se han de tener en cuenta las particularidades que puede derivarse de la inclusión en dicho tratamiento de operaciones[1] (artículo 4.2) que se implementen con componentes basados en soluciones tecnológicas específicas. Dichos componentes son implementaciones concretas de técnicas de tratamiento de datos que, por su proporcionalidad, efectos colaterales u otros aspectos, pueden incorporar elementos característicos que afecten tanto al cumplimiento del RGPD como a los riesgos que se derivan para los derechos y libertades de los interesados. Además, hay que recordar la necesidad de informar el interesado de la existencia de decisiones basadas en tratamientos automatizados, incluida la elaboración de perfiles (artículo 13.2.f)[2], además de proporcionar información significativa sobre la lógica aplicada, de la importancia y las consecuencias previstas de dicho tratamiento para el interesado.

El potencial impacto que los tratamientos basados en componentes IA podrían tener en los derechos y libertades de los ciudadanos pone de manifiesto la necesidad de establecer medidas de control efectivo, corrección, responsabilidad, rendición de cuentas, gestión del riesgo y transparencia relativas a los sistemas y tratamiento de los datos en los que se utilicen. Actualmente, los modelos de auditoría de tratamientos que incorporan componentes IA están en desarrollo. También están en desarrollo modelos holísticos que permitan implementar en la práctica el principio de rendición de cuentas (en inglés, accountability) durante todo el ciclo de vida del dato y que permitan depurar responsabilidades en las distintas fases de recolección y tratamiento de los datos personales. En este sentido este documento representa solamente una primera aproximación hacia la determinación de elementos básicos que las futuras auditorías

[1] Operaciones según el artículo 4.2 del RGPD son: "recogida, registro, organización, estructuración, conservación, adaptación o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma de habilitación de acceso, cotejo o interconexión, limitación, supresión o destrucción".

[2] Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. 2017. "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law* 7 (2):76-99.

podrían incorporar desde el punto de vista de la protección de datos y hacia la elaboración de futuros estándares[3]..

En el documento "Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial. Una introducción", publicado por la AEPD, se dedicaba un capítulo a la auditoría de tratamientos que incluyesen bien soluciones de IA o bien fases implementadas mediante componentes de IA, vista esta desde una perspectiva de la normativa de protección de datos personales. Se planteaba la auditoría como una de las posibles herramientas de evaluación en el marco de los tratamientos que utilicen IA y un instrumento dirigido a conseguir productos seguros, predecibles, controlables y cuya lógica interna se pueda en cierto modo explicar.

En este sentido, hay que tener en cuenta que las soluciones de inteligencia artificial, especialmente las que se conocen como de tipo "IA-débil" (Ver Anexo I "Definiciones"), se construyen y ejecutan utilizando componentes tradicionales de hardware y software. La "IA-débil", antes que un nuevo modelo de computación es un nuevo modelo de desarrollo de aplicaciones. La madurez de dicho proceso de desarrollo será crítica para garantizar la trazabilidad, explicabilidad y calidad del producto construido.

Los modelos de desarrollo tradicionales están soportados por buenas prácticas, ampliamente conocidas e implantadas, como la gestión del ciclo de vida de desarrollo software (SDLCM, del inglés Software Development Life Cycle Management), los modelos de madurez de las capacidades (CMM, del inglés Capacity Madurity Model) o los modelos de gestión del ciclo de vida de las aplicaciones (ALM, del inglés Application Life Cycle Management). Todos estos modelos establecen guías y recomendaciones para el desarrollo sistemático de productos en general y de productos software en particular. Sin embargo, aunque teniendo puntos en común, es necesario adaptar estos modelos para casos concretos, en particular para componentes cuyo ciclo de vida difieren de los modelos de desarrollo de sistemas tradicionales[4]. Por ejemplo, este es el caso que se deriva de las características inherentes al aprendizaje basado en datos[5] (Machine Learning en inglés). Además, el ritmo de producción hace que se integren fuentes de datos, software y hardware propios y de terceros y que estos proyectos de datos pueden utilizar técnicas estadísticas, de aprendizaje automático o procedimientos más avanzados de inteligencia artificial. El ciclo continuo de integración de estos tres elementos es lo que confiere la singularidad al desarrollo de productos de IA.

A pesar de estas peculiaridades, en el desarrollo de componentes IA aplican los principios básicos de análisis, diseño, desarrollo, verificación y validación que han de ajustarse a las características propias de esta tecnología. De ahí la importancia de abordar

[3] Véase por ejemplo los esfuerzos de estandarización que se están llevando a cabo desde el comité ISO/IEC JTC 1/SC 42, formado por International Electrotechnical Commission (IEC) e International Standard Association (ISO), y que lleva a cabo actividades de normalización en el área de la inteligencia artificial. Relevante al respecto las recomendaciones de ese comité a la estrategia de IA de la UE: "CEN-CENELEC response to the EC White Paper on AI, version 2020-06". Disponible en: https://www.cencenelec.eu/News/Policy_Opinions/PolicyOpinions/CEN-CLC%20Response%20to%20EC%20White%20Paper%20on%20AI.pdf

[4] Mientras que las aplicaciones software tradicionales son deterministas y están programadas para comportarse conforme a unas especificaciones y requisitos concretos, las aplicaciones basadas en aprendizaje automático son probabilísticas, aprenden de datos, en su mayor parte no estructurados, y necesitan ser entrenadas a lo largo de un número variable de iteraciones a lo largo de diferentes etapas.

[5] Rama Akkiraju, Vibha Sinha, Anbang Xu, Jalal Mahmud, Pritam Gundecha, Zhe Liu, Xiaotong Liu, John Schumacher . Characterizing machine learning process: A maturity framework. IBM Almaden Research Center, San José, California, USA, nov 2018. Disponible en: https://arxiv.org/ftp/arxiv/papers/1811/1811.04871.pdf

un modelo sistemático en el ciclo de vida del desarrollo de los componentes IA, de modo que su proceso de construcción pueda ser auditado con garantías de calidad.

Al realizar la evaluación del tratamiento éste ha de analizarse como un todo. Sin embargo, la complejidad de determinadas soluciones tecnológicas, como aquellas basadas en IA, recomienda el apoyarse en directrices específicas para gestionar los elementos singulares[6,7,8,9,10] que dichas tecnologías incorporan. Dichas directrices o recomendaciones han de integrarse en el conjunto de controles generales dirigidos a la evaluación de un tratamiento concreto, convirtiéndose en un subconjunto particular y necesario en el análisis global del este. Es decir, son recomendaciones de carácter general y transversales para todo tratamiento que incluya dicho componente tecnológico, pero no exhaustivas de cara a realizar la evaluación de un tratamiento concreto en todo su alcance y dimensiones.

El libro blanco de la Comisión Europea "*White Paper On Artificial Intelligence - A European approach to excellence and trust*"[11], publicado en febrero de 2020, manifiesta que "*Los centros de pruebas deben permitir la auditoría y la evaluación independientes de los sistemas de IA de acuerdo con los requisitos descritos anteriormente. La evaluación independiente aumentará la confianza y asegurará la objetividad. También podría facilitar el trabajo de las autoridades competentes pertinentes.*". El Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial de la Comisión Europea enumera, además, entre los principios para el desarrollo de una IA confiable (en inglés, *Trustworthy AI*) la preservación de la privacidad, la equidad, transparencia y la rendición de cuentas en la gobernanza de los datos, la solidez y seguridad técnica de los componentes de IA, el respeto del ambiente y del bienestar y de la autonomía de las personas[12].

En el caso de este documento, estaríamos tratando de una lista de objetivos de control y controles de auditoría que podrían incluirse como parte del conjunto de controles de auditoría de un tratamiento que implemente uno (o más de uno) componentes basados en IA. La lista tiene un carácter extensivo, de modo que no todos los objetivos y controles que se presentan en el documento tienen por qué aplicarse en todos los tratamientos que incluyan un componente IA, sino que dependerá del caso concreto. La lista de controles es una propuesta de referencia para que el auditor seleccione aquellos que sean aplicables al tratamiento concreto que se audite. La selección se realizará en función de distintos factores: si afectan al cumplimiento del RGPD, del tipo de componente IA utilizado, del tipo de tratamiento (p.ej. si se audita el tratamiento de desarrollo del componente o de si se audita le tratamiento que incluye un componente en explotación) y, sobre todo, del riesgo que representa para los derechos y libertades.

---

[6] Guía sobre el uso de las cookies. AEPD, 2020. Disponible en: https://www.aepd.es/sites/default/files/2020-07/guia-cookies.pdf

[7] Análisis de los flujos de información en Android. Herramientas para el cumplimiento de la responsabilidad proactiva. AEPD, 2019. Disponible en: https://www.aepd.es/sites/default/files/2019-09/estudio-flujos-informacion-android.pdf

[8] Orientaciones para prestadores de servicios de Cloud Computing. AEPD, 2018.Disponible en: https://www.aepd.es/sites/default/files/2019-09/guia-cloud-prestadores.pdf

[9] Código de buenas prácticas en proyectos de big data. AEPD, 2017. Disponible en: https://www.aepd.es/media/guias/guia-codigo-de-buenas-practicas-proyectos-de-big-data.pdf

[10] Orientaciones y garantías en los procedimientos de Anonimización de datos personales. AEPD, 2016. Disponible en: https://www.aepd.es/sites/default/files/2019-09/guia-orientaciones-procedimientos-anonimizacion.pdf

[11] Libro Blanco sobre la Inteligencia Artificial - un enfoque europeo orientado a la excelencia y la Confianza [en línea] COM(2020) 65 final, p.30 [Consulta: 29 de octubre del 2020]. Disponible en https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

[12] AI-HLEG. 2019. Policy and investment recommendations for trustworthy Artificial Intelligence. High-Level Expert Group on Artificial Intelligence. Disponible en: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

Por otro lado, los controles presentados no cubren otros aspectos del tratamiento que no estén directamente vinculados a la incorporación de un componente de IA y que se derivan de un análisis del tratamiento con carácter general.

En este sentido, el presente documento no describe cómo realizar la auditoría de protección de datos de todo un tratamiento que se apoye en una solución de IA, ni pretende cubrir objetivos más allá de los establecidos en el RGPD, como podrían ser aspectos éticos o de eficiencia. Este documento tampoco tiene por objeto describir los aspectos generales de la metodología de auditoría, ampliamente establecidos en la literatura, ni recomendar herramientas concretas que se pudieran utilizar en la ejecución de las auditorías, ya sean estas generales o específicas. Este documento se circunscribe a ofrecer las orientaciones metodológicas particulares y un listado de objetivos de control y controles específicos que podrían ser seleccionados para incluir en el proceso de auditoría de protección de datos de un tratamiento que incorpore componentes o soluciones de IA.

En este contexto, el presente documento está dirigido, principalmente, a responsables que han de auditar tratamientos que incluyan componentes basados en IA, de cara a ser una ayuda a la hora de garantizar y poder demostrar el cumplimiento de obligaciones y principios en materia de protección de datos a los que están sujetos; a encargados y desarrolladores que quieran ofrecer garantías sobre sus productos y servicios; a los Delegados de Protección de Datos encargados tanto de supervisar los tratamientos como de asesorar a los responsables y por último, a los equipos de auditores cuando se ocupen de la evaluación de dichos tratamientos.

II. METODOLOGÍA DE AUDITORÍA Y TRATAMIENTOS QUE INCORPORAN COMPONENTES DE IA

Desde la protección de datos, es importante recordar que la aproximación hacia la IA vendrá siempre de la perspectiva del uso de tecnologías emergentes cuyos riesgos para los derechos y las libertades de los interesados habrá que evaluar. Dicha evaluación llevará de forma natural a la adopción de medidas técnicas y organizativas que permitan minimizar dichos riesgos y maximizar los beneficios esperados del tratamiento de los datos personales. Desde esta aproximación, resulta evidente el valor de las metodologías ya en uso y de los estándares y certificaciones existentes.

Con independencia de que las diferentes metodologías de realización de auditorías existentes pueden diferir unas de otras en función del enfoque o perspectiva hacia el que estén focalizadas, se pueden encontrar múltiples referencias aplicables. Por ejemplo, los principios generales en la ISO 19011[13], los aspectos más específicos sobre el desarrollo software de sistemas de información en normas como ISO/IEC 15504 SPICE[14], metodologías como METRICAv3[15], estándares como CMMI[16] u otros marcos de referencia como COBIT[17], SOGP[18].

En este documento no se van a detallar los aspectos metodológicos de una auditoría ya que se encuentran ampliamente desarrollados en las anteriores reseñas. Sin embargo, existen características diferenciadoras que hay que tener en cuenta a la hora de planificar y ejecutar el proceso de una auditoría algorítmica de IA. Existen además numerosos esfuerzos de estandarización[19] [20] en el dominio de la inteligencia artificial que habrá que tomar en cuenta a la hora de desarrollar procesos de auditoría concretos que se ajusten a las necesidades de los tratamientos con componente de IA objeto de evaluación.

A. OBJETIVOS GENERALES DE LA AUDITORÍA DE UN COMPONENTE IA EN PD

Una auditoría de un componente de IA en protección de datos ha de ser un proceso sistemático, independiente y documentado para obtener evidencias y evaluarlas de manera objetiva con el fin de determinar el grado de cumplimiento de los criterios de auditoría que se hayan determinado y que, en el caso que nos ocupa, están en relación con los principios

[13] Directrices para la auditoría de sistemas de gestión. ISO 19011:2018. Disponible en: https://www.iso.org/standard/70017.html y https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma?c=N0060855

[14] ISO/IEC 15504-1:2004 Information technology – Process assessment – Part 1: Concepts and vocabulary. Disponible en: https://www.iso.org/standard/38932.html

[15] Métrica v3. Consejo Superior de Informática, 2001. Disponible en: https://administracionelectronica.gob.es/pae_Home/pae_Documentacion/pae_Metodolog/pae_Metrica_v3.html

[16] The Capability Maturity Model Integration (CMMI). Carnegie Mellon University (CMU). Disponible en: https://cmmiinstitute.com/cmmi/intro

[17] Control Objectives for Information and Related Technologies (COBIT). ISACA. Disponible en: https://www.isaca.org/resources/cobit

[18] Standard of Good Practice for Information Security 2020. Information Security Forum (ISF). Disponible en: https://www.securityforum.org/tool/standard-of-good-practice-for-information-security-2020/

[19] Véase al respecto el posicionamiento de European Telecommunications Standards Institute (ETSI) y de International Telecommunication Union (ITU) al respecto ("White Paper No. #34 Artificial Intelligence and future directions for ETSI". ETSI. 1st edition: June 2020. Disponible en: https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp34_Artificial_Intellignce_and_future_directions_for_ETSI.pdf. "Artificial Intelligence (AI) for Development Series Module on Setting the Stage for AI Governance: Interfaces, Infrastructures, and Institutions for Policymakers and Regulators" ITU. July 2018. Disponible en: https://www.itu.int/en/ITU-D/Conferences/GSR/Documents/GSR2018/documents/AISeries_GovernanceModule_GSR18.pdf), y de

[20] Al respecto es pertinente tener en cuenta la acción de los siguientes comités de estandarización: CTN27 (ISO/IEC 27037:2012); ISO 27050 electronic Discovery, CTN320, y el TC 46 SC11 (Record´s Management), que es el comité de estanadarización ISO sobre Inteligencia Artificial y Big Data.

de tratamiento y el resto de los requisitos establecidos por el RGPD y el resto de normativa de protección de datos aplicable.

El ciclo de vida de un componente IA se puede estructurar en varias etapas, como la concepción y análisis del componente, su desarrollo (incluyendo la investigación, selección, análisis y depuración de los conjuntos de datos, prototipado, diseño, entrenamiento, pruebas, implementación en software y/o hardware, integración como parte de un tratamiento y validación global), su explotación (incluyendo su mantenimiento evolutivo) o la retirada final del mismo. Las etapas anteriores pueden admitir iteraciones, dependiendo del modelo de desarrollo seleccionado.

Como en todo proceso de auditoría, antes de llevarla a cabo es necesario determinar su alcance. Este alcance puede cubrir todos los aspectos del ciclo de vida del desarrollo del componente y del tratamiento en el que se va a emplear. En este caso, estaríamos ante un alcance total de la auditoría. También es posible limitar el alcance de la auditoría a etapas y aspectos concretos del componente de IA. Por ejemplo, una auditoría limitada a los conjuntos de datos empleados, los aspectos de la metodología de desarrollo seguida, la implementación del diseño del componente en un elemento físico o librería software en un entorno concreto o la forma en que se produce la integración del componente como parte de un tratamiento que se despliega en un contexto determinado.

Dentro del vasto abanico de las auditorías de calidad, se pueden llevar a cabo tanto auditorias de proceso como de producto. El objetivo de la auditoría podrá garantizar el cumplimiento normativo con relación a la normativa de datos de carácter personal y así prevenir instancias de incumplimiento. De la misma forma, se podrán identificar, anticipar y corregir los posibles riesgos asociados a la utilización de datos personales en los componentes IA. Esto permite, a su vez, reforzar los mecanismos de responsabilidad proactiva implementados por el responsable del tratamiento de cara a demostrar el cumplimiento de sus obligaciones en materia de protección de datos. La auditoría también puede contribuir a realizar un análisis detallado del funcionamiento del componente IA permitiendo así la implementación de mecanismos de transparencia de cara al responsable que le permitan conocer y poder justificar las decisiones de diseño, desarrollo y/o selección del sistema.

B. CARACTERÍSTICAS SINGULARES DE LA METODOLOGÍA DE LA AUDITORÍA DE UN COMPONENTE IA EN PD

Como en toda auditoría, la primera decisión a adoptar es la definición de su objetivo y alcance, que podrá estar determinado por los intereses de la organización o por imposiciones externas (certificaciones, demostrar cumplimiento ante terceros u otras obligaciones normativas).

Los criterios que han de guiar una auditoría en protección de datos de un tratamiento en el que se incluya un componente IA son los principios relativos al tratamiento definidos por el Reglamento: licitud, lealtad y transparencia, limitación de la finalidad, minimización de datos, exactitud, limitación del plazo de conservación, integridad y confidencialidad, y responsabilidad proactiva (art. 5 del RGPD).

La selección de los objetivos de control y controles a tener en cuenta en el proceso de auditoría, la extensión de su análisis por parte del auditor y la formalidad que el auditor exija en la implementación de cada control dependerá, como en toda auditoría, del objetivo y alcance que se haya definido para esta, así como del análisis de riesgos realizado por el auditor. Como el objeto de la auditoría que cubre este documento es el de cumplimiento del

RGPD, el análisis de riesgos ha de realizarse circunscrito al ámbito de los derechos y libertades de las personas cuyos datos son tratados.

Además, hay que tener en cuenta el análisis de los riesgos de la auditoría en sí misma como proceso, con relación a cumplir los objetivos marcados en el contexto del componente IA concreto. El análisis de riesgos del proceso de auditoría está orientado a identificar aquellos aspectos que pueden afectar al logro de los objetivos definidos y que estarán vinculados con la planificación, los recursos disponibles, la selección del equipo auditor, el control de la información documentada, la disponibilidad y cooperación de las personas responsables, condiciones relativas tanto al mismo componente IA como al contexto en el que se inscribe, los procesos de seguimiento y revisión del programa de auditoría diseñado, cuestiones administrativas y legales o cualquier otra circunstancia.

Por lo tanto, como en todo documento de recomendaciones de auditoría, la enumeración en este documento de un conjunto de posibles controles no implica la obligación de aplicar sistemáticamente todos y cada uno de ellos, sino que hay que seleccionar, de forma racional, aquellos relevantes para cumplir con el objeto y el alcance definidos para la auditoría. Así, por ejemplo, en el caso de un componente que, a partir del análisis de los datos de entrada, pueda llegar a tomar decisiones que afecten de manera significativa a un individuo, negándole el acceso a servicios esenciales o restringiendo sus libertades, resulta evidente que el alcance de la auditoría y la exhaustividad del análisis de los controles propuestos ha de ser mayor que en un componente que, por ejemplo, se limite a la clasificación del correo electrónico en la bandeja de *spam*.

Para que los objetivos de control y controles detallados en el siguiente capítulo sean aplicables, se ha de partir de la premisa de que existe, o se pretende realizar, un tratamiento de datos personales en alguna de las etapas del ciclo de vida del componente IA, o que el tratamiento se dirige a perfilar al individuo o a ejecutar decisiones automáticas sobre personas físicas que tengan efectos jurídicos sobre ellas o les afecten significativamente. Esto podría implicar, en algunos casos, un análisis previo del grado de anonimización alcanzado para los datos utilizados en el tratamiento en general y por el componente en particular, el cálculo o estimación del posible riesgo de reidentificación que existe, o del cálculo del riesgo de pérdida de datos en la nube[21], entre otros.

Dado el tipo de auditoría que se está tratando, el equipo auditor ha de contar con personal con conocimientos tanto en la solución de IA objeto de auditoría como en la normativa de protección de datos. Además, en el caso de que en la organización del responsable del tratamiento esté nombrado un Delegado de Protección de Datos o DPD, este tendría que estar a disposición del equipo auditor de modo que pueda aclarar cualquier cuestión relativa a la finalidad, naturaleza, alcance y contexto del tratamiento que condicione el comportamiento del componente IA utilizado. Así mismo, y dependiente del modelo de desarrollo, podría ser aconsejable que formase parte del equipo auditor un científico de datos.

[21] La prevención de pérdida de datos en la nube (en inglés Cloud Data Loss Prevention o DLP) puede calcular cuatro métricas de riesgo de reidentificación: k-anonimato, l-diversidad, k-map, y δ-presencia. Un conjunto de datos es k-anónimo si los cuasi-identificadores de cada persona en el conjunto de datos son idénticos a por lo menos k - 1 otras personas también en el conjunto de datos. Un conjunto de datos tiene l-diversidad si, para cada conjunto de filas con cuasi-identificadores idénticos, hay al menos l valores distintos para cada atributo sensible. K-map calcula el riesgo de reidentificación comparando un determinado conjunto de datos de sujetos desidentificados con un conjunto de datos de reidentificación o "ataque" más grande. δ-presencia estima la probabilidad de que un usuario determinado de una población más grande esté presente en el conjunto de datos. Se utiliza cuando la pertenencia al conjunto de datos es en sí misma información sensible. Más información disponible en: https://cloud.google.com/dlp/docs/compute-risk-analysis

III. OBJETIVOS DE CONTROL Y CONTROLES

En este apartado se enumeran un conjunto de objetivos de control y de posibles controles para tener en cuenta cuando se realicen auditorías de componentes de inteligencia artificial incorporados a tratamientos de datos personales y/o que toman decisiones automatizadas que afecten a personas físicas.

Como se ha indicado anteriormente, la selección de los controles a auditar, la extensión de su análisis y la formalidad requerida en su implementación dependerá, como en toda auditoría, del objetivo y alcance definido para esta, así como del análisis de riesgos realizado. Por lo tanto, el auditor ha de seleccionar, del listado de controles propuesto, aquellos que se adecuen a la auditoría concreta y añadir aquellos que estime oportunos.

A. IDENTIFICACIÓN Y TRANSPARENCIA DEL COMPONENTE

Objetivo: Inventario del componente IA auditado

En cumplimiento de la responsabilidad proactiva[22] debe existir trazabilidad, y para ello, una correcta identificación del componente IA parte del tratamiento auditado.

Controles:

- El componente IA está identificado en la documentación con un nombre o código, la identificación de la versión[23] y la fecha de creación.
- Tanto el código como cualquier archivo adicional que defina la versión dispondrá de una firma digital de todo el conjunto que garantice su integridad.
- Existe y está documentado un histórico de versiones de la evolución del componente IA utilizado, incluyendo los parámetros usados en el entrenamiento del componente y todo aquello que asegure la trazabilidad de la evolución/cambios en el componente.

Objetivo: Identificación de responsabilidades

En cumplimiento de la responsabilidad proactiva debe existir una identificación clara de los roles con relación al tratamiento auditado que incluye el componente IA y las responsabilidades de las partes implicadas[24] [25].

Controles:

- Datos identificativos y de contacto de la/s persona/s o institución/instituciones responsables de las etapas del ciclo de vida del componente IA bajo auditoría y/o, corresponsables, representantes del responsable y de los encargados.

[22] Artículo 5.2 del RGPD – Principios relativos al tratamiento. Responsabilidad proactiva. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

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

[24] Capítulo IV del RGPD – Responsable del tratamiento y encargado del tratamiento. . Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3085-1-1

[25] El proceso de depuración de responsabilidades incluye toda la cadena de suministro del sistema en que se implementa el componente de IA. En los procedimientos actuales de integración continua de software y de hardware estos ciclos llevan asociadas cadenas de dependencia donde hay que lograr una adecuada convergencia entre confianza de los usuarios y fiabilidad de las maquinas Véase por ejemplo el libro de Olav Lysne "The Huawei and Snowden Questions: Can Electronic Equipment from Untrusted Vendors be Verified? Can an Untrusted Vendor Build Trust into Electronic Equipment?". 2018. Springer Nature

- Especificación en los contratos asociados a las etapas de tratamiento bajo auditoría del reparto de responsabilidades desde el punto de vista de protección de datos personales.
- Inscripción en el Registro de Actividades de Tratamiento de los responsables respectivos, y/o los encargados, del tratamiento de datos personales bajo auditoría.
- Determinación de si existe obligación de disponer de un Delegado de Protección de Datos, y en caso afirmativo, identificación de este y comunicación ante la Autoridad de Control.

Objetivo: Transparencia

En cumplimiento del principio de transparencia[26] y de la obligación de proporcionar la información relativa al tratamiento a los interesados[27], tanto el origen de los datos como las propiedades y la lógica del componente IA es accesible, comprensible y puede ser explicada.

Controles:

- El origen de los datos está documentado y existe un mecanismo para informar.
- Las características de los datos usados para entrenar al componente IA están identificadas, documentadas y adecuadamente justificadas.
- Teniendo en cuenta criterios de eficiencia, calidad y precisión del componente IA, se ha elegido el modelo más adecuado (usando criterios de simplicidad e inteligibilidad), entre varios componentes concurrentes[28], y desde el punto de vista de su codificación para facilitar la legibilidad, comprensión de su lógica, la coherencia interna y la explicabilidad[29].
- La información sobre los metadatos[30] del componente IA, su lógica y las consecuencias que pueden derivarse de su empleo están accesibles para las partes interesadas junto con los medios o mecanismos disponibles para ejercer sus derechos en caso de objeción respecto de los resultados.
- Existe la documentación suficiente para comprender la lógica del componente IA utilizado y realizar la trazabilidad de su comportamiento respecto a cada conjunto de datos de entrada, el uso que hace de estos y de los datos intermedios, y cómo entrega los datos de salida.
- Ante un comportamiento erróneo por parte del componente IA que pueda ocasionar perjuicios a los interesados, se han previsto mecanismos para

[26] Artículo 5.1.a y Capítulo III – Sección 1 del RGPD - Principios relativos al tratamiento. Licitud, lealtad y transparencia. Disponible en: (https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1) y Transparencia y modalidades (https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2235-1-1) respectivamente.

[27] Artículos 13.2.f y 14.2.g del Capítulo III - Sección 2 del RGPD Información y acceso a los datos personales. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2307-1-1

[28] Habrá que tomar en consideración el estado actual del desarrollo del componente de IA y contar con el asesoramiento de expertos e investigadores en inteligencia artificial (aprendizaje automático profundo, procesamiento de lenguaje natural, etc.). Sobre aprendizaje automático profundo véase por ejemplo la revisión de Xie, N., Ras, G., van Gerven, M. and Doran, D., 2020, "Explainable deep learning: A field guide for the uninitiated." *arXiv preprint arXiv:2004.14545.*

[29] En los últimos tiempos ha habido una creciente preocupación por comprender y explicar cómo los modelos de Inteligencia Artificial toman decisiones a partir de los datos capturados en su entorno de aplicación, particularmente en aquellos donde dichas decisiones tienen implicaciones para el ser humano (e.g. diagnóstico médico). Subyacen varios motivos para esta creciente preocupación.http://blogs.tecnalia.com/inspiring-blog/2019/11/14/explicabilidad-transparencia-trazabilidad-equidad-no-precision-uso-responsable-la-inteligencia-artificial/

[30] Los metadatos del componente IA son los parámetros utilizados en los procesos de aprendizaje.

minimizar esos perjuicios dar soporte a la comunicación a las partes interesadas y facilitar la comunicación entre todas las partes involucradas en el proceso.

B. PROPÓSITO DEL COMPONENTE IA

Objetivo: Identificación de las finalidades y usos previstos

En cumplimiento del principio de limitación de finalidad[31], los fines para los que son empleados los datos procesados por y para el componente IA, han de ser determinados, explícitos y legítimos sin ser utilizados de manera incompatibles con aquellos.

Controles:

- Está documentado el objetivo que se persigue con el uso del componente IA, tanto en términos cuantitativos como cualitativos, con una descripción clara de lo que se pretende conseguir mediante su empleo en el marco del tratamiento.
- Está documentada la relación entre el objetivo que se persigue con el uso del componente de IA en un determinado tratamiento y las condiciones que garantizan la licitud de dicho tratamiento.
- Están identificadas las dinámicas, actividades y/o procesos en el marco de la organización en la que se integra la etapa del ciclo de vida del componente IA bajo auditoría, delimitando, en la medida de lo posible, el contexto de su uso.
- Los usuarios potenciales del componente IA están categorizados.
- En su caso, se han descrito otros posibles usos y usuarios secundarios[32] junto con la base de legitimación que justifica su utilización;

Objetivo: Identificación del contexto de uso del componente IA

En cumplimiento de la obligación de analizar el contexto del tratamiento en el que se integra el componente IA[33] se han de conocer las circunstancias en las que tiene lugar el tratamiento así como otros factores que puedan condicionar las expectativas de las partes concernidas y que puedan tener un impacto sobre los derechos de los interesados. El análisis de esas circunstancias ayudará a poder determinar las medidas técnicas y organizativas más adecuadas de cara a garantizar el cumplimiento de las obligaciones normativas.

Controles:

- Están documentados los contextos jurídico, social, económico, organizacional, técnico, científico, o de cualquier otra clase que esté relacionado con la inclusión del componente IA en el tratamiento de datos personales.
- Está definida en la estructura organizacional y/o contractual entre las partes y así el reparto de tareas y responsabilidades.
- Están descritos los factores condicionantes de la efectividad del componente, incluyendo las garantías jurídicas, las normas aplicables, los recursos organizativos y técnicos, los datos disponibles y las dinámicas internas que ha

[31] Artículo 5.1.b del RGPD – Principios relativos al tratamiento. Limitación de la finalidad. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[32] 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.

[33] Artículo 24.1 del RGPD – Responsabilidad del responsable del tratamiento. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3096-1-1

de incorporar un tratamiento de datos personales para la inclusión, con garantías, del componente IA bajo auditoría.

- Están definidos los requisitos de los operadores humanos que tendrán misión de supervisar e interpretar la operación del componente IA.
- En su caso, está documentada la interacción del componente IA con otros componentes, sistemas o aplicaciones, propias o de terceros y el reparto de responsabilidades de mantenimiento, actualización y minimización de los problemas de privacidad del sistema..
- Están definidos los baremos o umbrales para interpretar y utilizar los resultados ofrecidos por el componente IA utilizado.
- Están identificados y descritos aquellos contextos para los que no está recomendado incluir el componente IA en un tratamiento al no poder cumplir con su objeto o propiedades, o por tener el componente un nivel de fiabilidad y/o exactitud inadecuado respecto a la relevancia que podría tener el tratamiento en el interesado[34].

Objetivo: Análisis de la proporcionalidad y necesidad

Cuando el componente IA se audite en el marco de un tratamiento en el que sea obligatorio realizar una evaluación de impacto relativa a la protección de datos, se ha de analizar la necesidad y proporcionalidad[35] que supone su empleo respecto de la finalidad perseguida.

Controles:

- Existe una evaluación del empleo del componente IA en el marco del tratamiento frente a otras posibles opciones con relación a los derechos y libertades de los interesados.
- En el caso de nuevos desarrollos, se ha realizado un análisis comparado de la eficacia y la adecuación de los resultados del componente IA frente a otros componentes más probados, que utilizan criterios más estrictos de minimización o que presentan menos riesgos para los derechos y libertades de las personas, en particular aquellos que hacen un uso menos intensivo de categorías especiales de datos.
- En el caso de abordar un problema nuevo, existen, se han documentado y se pueden justificar las motivaciones y argumentos que conducen a abordar este problema a través del empleo de un componente IA.
- En el caso de abordar un problema conocido, existen, se han documentado y se pueden justificar los motivos que han conducido a un cambio en el esquema de funcionamiento anterior, describiendo, en su caso, los nuevos objetivos que se persiguen mediante el empleo del componente IA en el marco del tratamiento.
- Existe un análisis y gestión del riesgo para los derechos y libertades de los interesados que introduce en el tratamiento el procesamiento de los datos mediante el componente IA.

[34] Es importante tomar en cuenta a la hora de realizar dicho análisis de contexto la repercusión que tendrá el tratamiento en cuestión sobre la vida del interesado, especialmente para oportunidades presentes y futuras.

[35] Artículo 35.7.b del RGPD Evaluación de impacto relativa a la protección de datos. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3606-1-1

Objetivo: Determinación de los destinatarios de los datos

En cumplimiento de las obligaciones derivadas de la atención de los derechos de los interesados[36] y en especial los relativos a la transparencia y la información facilitada[37] se han de identificar los destinatarios o categorías de destinatarios de los datos personales procesados por el componente IA, incluidos los destinatarios en terceros países u organizaciones internacionales.

Controles:

- Están identificadas las obligaciones de información a los interesados con relación al tratamiento de datos derivado de la inclusión del componente IA.
- Dichas obligaciones se identifican tanto para los datos obtenidos directamente de los interesados como cuando no se obtienen directamente de ellos y proceden de otras fuentes.
- En la determinación de estas obligaciones están identificados los destinatarios o las categorías de destinatarios a quienes se comunicaron o se comunicarán los datos personales que trata el componente IA, incluidos los destinatarios en terceros países u organizaciones internacionales, tanto en el caso en que los datos se obtienen directamente de ellos o en el caso en que proceden de otras fuentes de información.
- En la determinación de estas obligaciones están identificadas las intenciones del responsable de transferir datos personales a un destinatario en un tercer país u organización internacional y la existencia o ausencia de una decisión de adecuación de la Comisión[38]. En el caso de transferencias mediante garantías adecuadas, las basadas en la aplicación de normas corporativas vinculantes o las que respondan a las excepciones a las que se refiere el artículo 49, apartado 1, párrafo segundo, se facilita referencia a dichas garantías y a los medios para obtener una copia de ellas o al hecho de que se hayan prestado[39].
- Los destinatarios de los datos, incluidos los de terceros países u organizaciones internacionales, aparecen identificados en la actividad o actividades del Registro de Actividades de Tratamiento en las que se inscribe el uso del componente IA.

Objetivo: Limitación de la conservación de datos

En cumplimiento del principio de limitación del plazo de conservación de los datos[40], y salvo las excepciones previstas, los datos utilizados por el componente IA, ya sean de entrenamiento o los generados por él, se mantendrán durante no más tiempo del necesario para los fines que se pretenden alcanzar[41].

---

[36] Capítulo III del RGPD – Derechos del interesado. https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2224-1-1

[37] Artículos 13.1.e y 14.1.e del RGPD - Información que deberá facilitarse cuando los datos personales se obtengan del interesado (https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2317-1-1) e información que deberá facilitarse cuando los datos personales no se hayan obtenido del interesado. Disponible en: (https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2418-1-1), respectivamente.

[38] European Commision, How the EU determines if a non-EU country has an adequate level of data protection. Disponible en: https://ec.europa.eu/info/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en

[39] AEPD. Transferencias internacionales. Disponible en: https://www.aepd.es/es/derechos-y-deberes/cumple-tus-deberes/medidas-de-cumplimiento/transferencias-internacionales

[40] Artículo 5.1.e del RGPD – Principios relativos al tratamiento. Limitación del plazo de conservación. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[41] El marco del tratamiento en el que se está realizando la auditoría podría ser el de desarrollo del componente o, en otros casos, el tratamiento podría ser aquel en el que se ha incorporado dicho componente

Controles:

- Están identificadas las bases legitimadoras para la conservación de los datos personales utilizados por el componente IA por un periodo de tiempo más allá del acotado a los fines del tratamiento, en particular, si está relacionado con fines compatibles o responde a alguna de las excepciones previstas por la normativa[42] (fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos).
- Están determinadas y justificadas las etapas de ciclo de vida del componente IA en las que es necesario conservar los datos personales tratados.
- Están definidos los criterios y las medidas técnicas y organizativas apropiadas para la conservación de los datos personales[43].
- Están determinados los plazos previstos para la supresión de los datos personales conservados.
- Está definida una política de conservación de una muestra de datos de entrenamiento con el objeto de auditar el componente de IA, en la que se contemplen los riesgos mínimos o asumibles para los interesados.
- Existen procedimientos para verificar la aplicación de criterios, medidas y plazos de conservación.
- Está definido un procedimiento de revisión del análisis de la necesidad y proporcionalidad de la conservación de los datos para aquellos casos en los que se haya detectado un patrón de conservación excesivo, ya sea en plazo o en cantidad.
- Está definida una política de conservación de los datos personales incluidos en los ficheros de registro de actividad del componente IA y se aplican estrategias de privacidad (minimización, ocultación, separación o abstracción de datos) para su explotación.

Objetivo: Análisis de las categorías de interesados

Cuando sea obligatorio realizar una evaluación de impacto relativa a la protección de datos[44] en el marco del tratamiento en el que se inscribe el componente IA, se ha de identificar las categorías de interesados a los que afecta el tratamiento y valorar la procedencia de involucrarlos (o a sus representantes) en el proceso de evaluación[45].

Controles:

- Está identificado las categorías de interesados a los que afecta el desarrollo del componente IA y su uso en el marco del tratamiento previsto.
- Están identificadas las consecuencias a corto y largo plazo que la implementación del componente IA puede suponer a las categorías de interesados.

[42] Artículo 89.1 del RGPD - Garantías y excepciones aplicables al tratamiento con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e6569-1-1

[43] En particular, las medidas de minimización, ofuscación y/o seudonimización adoptadas.

[44] Artículo 35.9 del RGPD – Evaluación de impacto relativa a la protección de datos. Opinión de los interesados. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3606-1-1

[45] Estudios recientes proponen habilitar mecanismos de contestación o apelación que permitan al interesado, o a grupos en representación de los mismos, de intervenir y cuestionar la lógica o el resultado del componente de IA en el caso de procesos de toma de decisiones automatizadas. Véase por ejemplo el trabajo de Kluttz, Kohli, y Mulligan de 2018 "Contestability and Professionals: From Explanations to Engagement with Algorithmic Systems." Disponible en *SSRN 3311894*.

- Están definidos los procedimientos necesarios para analizar el contexto social en el que se enmarca el uso del componente y recabar información a través de personas, grupos u organizaciones afectadas por el mismo con el objetivo de conocer sus niveles de satisfacción, posturas, inquietudes e incertidumbres con respecto a la utilización de esta técnica en el marco del tratamiento de sus datos.

C. FUNDAMENTOS DEL COMPONENTE IA

Objetivo: Identificación de la política de desarrollo del componente IA

La política interna de desarrollo de sistemas, en particular de desarrollo de componentes de IA, ha de ser coherente con la política de protección de datos de la organización[46]. Será necesario detallar y complementar dicha política en los puntos específicos que sean necesarios, además de estar alineada con el RGPD, la LOPDGDD, y otra normativa sectorial que le sea de aplicación.

Controles:

- Los documentos con las políticas de desarrollo de productos y sistemas tienen en consideración la política de protección de datos.
- Existe un proceso de revisión y un control de versiones de las políticas.

Objetivo: Implicación del DPD

En cumplimiento de la posición y funciones atribuidas al Delegado de Protección de Datos[47] se han definido los procedimientos internos de comunicación y reparto de responsabilidades que permiten que su figura preste el debido asesoramiento y pueda participar activamente en la selección, diseño y/o desarrollo del componente IA en el que se apoya el tratamiento de los datos personales.

Controles:

- El Delegado de Protección de Datos reúne las cualidades profesionales necesarias y, en particular, los conocimientos especializados en materia de derecho, técnicos y la práctica en materia de protección de datos adecuados para el proyecto.
- El delegado de protección de datos cuenta con la ayuda y el asesoramiento de expertos en materias específicas relativas al componente IA objeto de auditoría.
- Están definidos procedimientos internos dentro de la organización para la correcta comunicación entre el Delegado de Protección de Datos y las personas a cargo de aquellos proyectos con implicaciones en el tratamiento de datos personales para que se pueda contar con su asesoramiento, en particular, durante el desarrollo de la evaluación de impacto de protección de datos de aquellos tratamientos que hacen uso de componentes de IA.
- El Delegado de Protección de Datos ha participado en las etapas objeto de auditoría, se ha respectado su independencia de juicio dentro de organización y sus obligaciones a cooperar con las agencias de supervisión y sus opiniones y consideraciones han sido tenidas en cuenta.

[46] Artículo 24.1 del RGPD – Responsabilidad del responsable del tratamiento. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3106-1-1

[47] Sección 4 del Capítulo IV del RGPD (artículos 37,38 y 39) – Delegado de protección de datos. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3782-1-1

Objetivo: Adecuación de los modelos teóricos base

El tratamiento, para que pueda ser considerado leal[48], ha de ser idóneo con relación al propósito del mismo que haya sido declarado.

Controles:

- Se ha realizado un estudio y análisis sobre el marco teórico y experiencias previas similares sobre las que se fundamenta el desarrollo del componente IA.
- Se han determinado, argumentado y documentado de forma precisa las ideas de base e hipótesis que se toman en consideración para la creación y desarrollo del modelo[49].
- Está definido un procedimiento de revisión crítica y contrastada de los razonamientos derivados de la aceptación de hipótesis importantes para el desarrollo del componente IA (por ejemplo, examinar cuáles son los argumentos tras una relación causal que modela un algoritmo, como la selección de variables que definen un fenómeno).
- Se ha realizado un análisis cuidadoso de cara a establecer presunciones adecuadas sobre las posibles variables *proxy* que intervienen en el componente IA.

Objetivo: Adecuación del marco metodológico

El tratamiento, para que pueda ser considerado leal[48], ha de ser idóneo con relación al propósito del mismo que haya sido declarado.

Controles:

- Debe estar documentado el marco metodológico de definición del modelo y creación del componente IA en las etapas objeto de auditoría como, por ejemplo, la forma de seleccionar, recoger y preparar los datos de entrenamiento del componente, realizar su etiquetado, construir el modelo, utilización de los datos intermedios generados, seleccionar el subconjunto de datos de test/validación o cómo medir los desajustes del modelo para mejorarlo.
- Está determinado, en función del análisis del problema a resolver y de manera justificada, el modelo de desarrollo a utilizar (p.ej.: supervisado, no supervisado u otros) y en el caso de los supervisados, el modelo de supervisión del aprendizaje del algoritmo, el grado de supervisión y su fundamento.
- Se han seleccionado y definido las métricas con respecto de las cuales medir el comportamiento del componente IA.
- Existe un procedimiento de registro y seguimiento de las desviaciones en el comportamiento del componente IA respecto de las métricas definidas que permite realizar una monitorización de aquellas circunstancias que pueden derivar en un comportamiento erróneo o sesgado del componente.

Objetivo: Identificación de la arquitectura básica del componente

En cumplimiento de la responsabilidad proactiva está documentado el desarrollo del componente IA de forma que es posible comprender la parte relativa a su implementación,

[48] Artículo 5.1.a del RGPD – Principios relativos al tratamiento. Licitud, lealtad y transparencia. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[49] 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.

su contexto de funcionamiento y las interrelaciones que mantiene con otros elementos integrantes del tratamiento.

Controles:

- En la fase de análisis del proyecto de desarrollo del componente IA, se han incluido, como parte del catálogo de requisitos, aquellos específicos para garantizar la privacidad y la protección de los datos personales.
- En la programación de los componentes de IA se han seguido y documentado los principios[50], códigos y buenas prácticas de codificación[51,52,53] utilizados para garantizar que el código sea legible, seguro, fácil de mantener y robusto.
- Está identificada y documentada la arquitectura básica del componente IA, incluyendo información sobre la técnica de aprendizaje automático utilizada, el tipo o tipos de algoritmos probados y, en su caso, descartados en la fase de aprendizaje y entrenamiento, y otros datos sobre el funcionamiento del componente como la función de pérdida o función de coste del modelo[54].
- Existe y funciona un procedimiento sistemático de documentación de la implementación del componente que garantiza el registro y posterior adquisición de toda la información necesaria para identificarlo, así como a sus elementos y a su entorno, comprender lo qué hace y por qué lo hace y poder contrastar la calidad y legibilidad del código de cara a su auditoría: descripción del/los lenguaje/s de programación utilizados, versión más reciente del código, código comentado, paquetes y librerías necesarios para su lectura, interfaces con otros componentes, en su caso, APIs[55] empleadas, documentos de utilidad como especificación de requisitos, análisis funcional, análisis orgánico, manuales, etc.
- En el caso de imposibilidad de acceder al código del componente IA, se ha aplicado un proceso de ingeniería inversa u otro método alternativo como el uso de pruebas de conocimiento cero (ZKP por sus siglas en inglés *Zero Knowledge Proof*) que, aún sin tener acceso al código, permitan profundizar en el funcionamiento del componente y determinar la lógica de las reglas aplicadas de cara a detectar incoherencias, manipulaciones directas y subestimación o sobreestimación de las variables utilizadas en el componente original.

[50] What Are The Best Software Engineering Principles?. Luminousmen Blog, 2020. Disponible en: https://luminousmen.com/post/what-are-the-best-engineering-principles

[51] Clean Code: A Handbook of Agile Software Craftsmanship, Robert C. Martin, Pearson 2008

[52] Resumen Clean Code. Samuel Casanova. Disponible en: https://samuelcasanova.com/2016/09/resumen-clean-code

[53] Clean Architecture: A Craftsman's Guide to Software Structure and Design, Robert C. Martin, Pearson 2017

[54] Una función de pérdida J(θ) mide el nivel de insatisfacción respecto de las predicciones del modelo con respecto a una respuesta correcta y utilizando ciertos valores de θ. Existen varias funciones de pérdida como el error cuadrático medio o entropía cruzada y la selección de uno de ellos depende de varios factores como el algoritmo seleccionado o el nivel de confianza deseado, pero principalmente depende del objetivo perseguido con el modelo. Como el propósito de entrenar un modelo es generar predicciones que estén lo más cerca posible de la respuesta correcta y minimizar el error o insatisfacción, el objetivo es encontrar los valores óptimos de θ que minimicen el resultado de la función de pérdida. Esto es lo que se conoce como optimización y, al igual que con la función de pérdida, existen varios métodos de optimización que impactan directamente sobre el rendimiento del modelo y el tiempo de entrenamiento. Uno de los métodos más utilizados es el método del gradiente descendente https://www.iartificial.net/gradiente-descendiente-para-aprendizaje-automatico/

[55] API: qué es y para qué sirve. XATACA, 2019. Disponible en: https://www.xataka.com/basics/api-que-sirve

D. GESTIÓN DE LOS DATOS

Objetivo: Aseguramiento de la calidad de los datos

En cumplimiento de los principios relativos al tratamiento[56] los datos personales tratados han de ser exactos y actualizados en relación con los fines para los que son tratados.

Controles:

- Existe un procedimiento documentado para gestionar y garantizar una adecuada gobernanza de los datos, que permita verificar y aportar garantías de la exactitud, integridad, fiabilidad, veracidad, actualización y adecuación del conjunto de datos utilizado en entrenamiento y/o prueba y/o explotación.
- Existen mecanismos de supervisión de los procesos de recopilación, tratamiento, conservación y utilización de los datos.
- Se ha realizado un análisis previo y una cuantificación de la muestra utilizada para el entrenamiento del modelo y se ha verificado que la adecuación del tamaño muestral, así como de la frecuencia y distribución de cada una de las variables, su intersección o grupos relevantes para el estudio, es el adecuado en relación con los parámetros definidos o con respecto a la realidad.
- Se han realizado análisis, tanto al inicio como en cada una de las iteraciones de las que se compone el proceso global de aprendizaje, de la muestra utilizada para el entrenamiento del modelo y se ha verificado que la representatividad del conjunto final de datos con relación a la población del contexto al que se orienta el componente IA y a los grupos definidos en el mismo es el adecuado.
- Se ha verificado que la distribución de las variables es adecuada y que el componente no es especialmente sensible o ignora alguna de ellas.
- Existen procedimientos para analizar, medir y detectar posibles desequilibrios entre la cantidad de datos que el componente recoge sobre una determinada variable con respecto a otra y que pueden dar lugar a desviaciones en su comportamiento.
- Se ha realizado 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 componente.
- Se ha realizado un análisis del tamaño muestral para la conservación de datos con propósito de auditoría.

Objetivo: Determinación del origen de las fuentes de datos

En cumplimiento de los principios relativos al tratamiento[57], los datos personales han de ser tratados con licitud, lealtad y transparencia y fines determinados, explícitos y legítimos (principio de limitación de la finalidad) con prohibición de tratar categorías especiales de datos[58] salvo en las circunstancias y excepciones previstas.

[56] Artículo 5.1.d del RGPD – Principios relativos al tratamiento. Principio de exactitud. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1863-1-1

[57] Artículo 5 del RGPD. Principios relativos al tratamiento. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1863-1-1

[58] Artículo 9 del RGPD – Tratamiento de categorías especiales de datos personales. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2114-1-1

Controles:

- El contexto de origen y las fuentes de datos utilizados para el entrenamiento y validación del modelo deben estar identificados.
- Se ha documentado y justificado el proceso de elección de las fuentes de datos utilizadas para el entrenamiento del componente.
- Existe una base de legitimación para el uso de los datos personales en las distintas etapas del ciclo de vida del componente IA.
- Existe una justificación para la recogida y el empleo de datos personales, que no son necesarios en la etapa de entrenamiento, para poder realizar una comprobación del comportamiento del modelo en las etapas posteriores de verificación y validación del componente[59].
- Si se realiza un tratamiento de datos personales sensibles, se ha evaluado la necesidad de su uso y existe una circunstancia que justifica levantar la prohibición general de su tratamiento.

Objetivo: Preparación de los datos personales

En cumplimiento de los principios relativos al tratamiento[60], los datos personales han de ser tratados aplicando el principio de minimización.

Controles:

- Los criterios para realizar la depuración previa de los conjuntos de datos originales y aquellas otras que se identifiquen como necesarias a lo largo de las diferentes iteraciones en el proceso de entrenamiento del componente, están identificados y documentados.
- Las técnicas y buenas prácticas de limpieza de datos[61] utilizadas en el proceso de depuración están argumentadas y documentadas.
- La clasificación de las variables define tipos claramente distinguibles e identificables.
- Está documentada la estructura y propiedades del conjunto de datos tratados, en número de sujetos y extensión de datos utilizados.
- Se ha realizado una categorización previa de los datos utilizados, organizándolos en datos no personales y personales, identificando, en el caso de estos últimos, qué campos constituyen identificadores, cuasi-identificadores y categorías especiales de datos.
- Se han determinado las variables relevantes para el modelo, identificando las variables asociadas a categorías especiales de datos y las variables *proxy*, incluyendo la información necesaria para su interpretación.
- Se han determinado y aplicado los criterios de minimización de los datos aplicables a las diferentes etapas del componente IA haciendo uso de aquellas estrategias de ocultación, separación, abstracción, anonimización y seudonimización de los datos que sean aplicables y conduzcan a maximizar la privacidad en el funcionamiento del componente.

[59] Información adicional asociada a los sujetos para un posterior análisis de si hay problemas de exactitud o sesgos asociados por características que no deberían estar asociadas al proceso de decisión. Por ejemplo, discriminación por sexo, racial, origen social, etc.

[60] Artículo 5 del RGPD – Principios relativos al tratamiento. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2114-1-1

[61] La limpieza de datos o *datacleaning* consiste en una serie de técnicas que, aplicadas sobre los registros de *dataset* o conjunto de datos, permiten su depuración reduciendo duplicaciones e inconsistencias. Más información disponible en: https://elitedatascience.com/data-cleaning

- Las bases de datos utilizadas tienen asociado un diccionario de datos que permite su análisis y comprensión.
- Están implementadas estrategias de segregación y disociación sobre la información adicional que no es necesaria para el entrenamiento pero que será necesaria en los procesos de verificación y validación del comportamiento del modelo para analizar correlaciones entre variables, comprobar el nivel de precisión del componente IA respecto a determinados atributos o verificar que no introduce sesgos[62].
- En la evaluación y selección de los datos se ha contado con la participación de un experto en técnicas de modelado y ciencia de datos encargado de entender los procesos complejos de la realidad que se intentan modelar, y de analizar e interpretar los datos utilizados por el componente.
- Se ha realizado un preprocesamiento previo y depuración de los datos utilizados para el entrenamiento y validación del componente IA, detectando las posibles anomalías que precisan un tratamiento previo (valores límite, registros incompletos, etc.) y convirtiendo las fuentes de datos heterogéneas a un único formato homogéneo.
- Se han introducido las modificaciones necesarias en el formato de los datos de entrada, si este no es adecuado en relación con el funcionamiento del componente IA o porque no representa la realidad que refleja[63].
- En su caso, se ha realizado análisis del grado de anonimización de los datos y del posible riesgo de reidentificación.
- En su caso, si se han usado técnicas de imputación de datos para completar la información del conjunto de datos, se han documentado los procedimientos y algoritmos usados para dicha imputación.

Objetivo: Control del sesgo

En cumplimiento de los principios relativos al tratamiento[64] los datos personales tratados han de ser exactos y actualizados con relación a los fines para los que son tratados.

Controles:

- Se han definido procedimientos para identificar y eliminar, o al menos limitar, los sesgos en los datos utilizados para entrenar el modelo.
- Se ha verificado que en los datos de entrenamiento usados como entrada al modelo no existen sesgos históricos previos y que, en caso contrario, o bien se ha optado por otra fuente de datos de entrenamiento que no los contenga, o bien se ha realizado una limpieza y depuración adecuadas para su normalización.
- Se han adoptado medidas para evaluar la necesidad de disponer de datos adicionales de cara a mejorar la precisión o eliminar posibles sesgos.
- Se han implementado mecanismos de supervisión humana para controlar y asegurar la ausencia de sesgos en los resultados.

---

[62] Por ejemplo, para controlar que un componente no discrimina por razones de género, aunque la variable género no se utilice durante el entrenamiento del componente debe haberse recogido información sobre el género de las personas que componen la base de datos para verificar si el componente se comporta de uno u otro modo en función del valor

[63] Por ejemplo, en el caso de un componente de procesamiento del lenguaje natural si la forma de funcionamiento de este no tiene la capacidad de adaptarse a cambios en las palabras que componen los textos de entrada, es probable que no se comporte de la manera deseada. En este caso se podrá optar por una forma de entrada de los datos que sea más ordenada y esquematizable o bien adaptar el funcionamiento del componente al formato de los datos de entrada que recibe.

[64] Artículo 5.1.d del RGPD – Principios relativos al tratamiento. Principio de exactitud. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1863-1-1

- Se han implementado mecanismos que permitan a los interesados solicitar la intervención humana, expresar su punto de vista e impugnar los resultados derivados del empleo de algoritmos automatizados en la toma de decisiones.

E. VERIFICACIÓN Y VALIDACIÓN

Objetivo: Adecuación del proceso de verificación y validación del componente IA

En cumplimiento de los principios relativos al tratamiento[65], en particular del principio de responsabilidad proactiva[66], se ha de poder demostrar que la metodología empleada en la incorporación del componente IA en el tratamiento, o en su desarrollo, cumple con los principios relativos al tratamiento y el resto de obligaciones impuestas por la normativa en materia de protección de datos.

Controles:

- Están documentados el proceso de verificación y validación, las técnicas empleadas, el conjunto de pruebas y comprobaciones realizadas, los resultados obtenidos y las acciones propuestas.
- Está establecida o se ha seguido una guía, norma o estándar para realizar un procedimiento sistemático de verificación y validación del componente IA y de su comportamiento una vez integrado en el tratamiento al que da soporte.
- Están implementados los mecanismos de control y supervisión necesarios para garantizar que el componente IA cumple los objetivos y propósitos previstos con eficacia.
- Están definidas y justificadas las métricas y criterios respecto a los cuales se realizará las comprobaciones en el proceso de verificación y validación.
- Está definida una estrategia de pruebas y, asociada a ella, existe un plan de pruebas completo para evaluar la corrección del componente IA tanto desde el punto de vista estructural como funcional.
- El personal involucrado en las tareas de verificación y validación del componente IA está cualificado para realizar las comprobaciones necesarias de cara asegurar que el componente se ha construido correctamente y se comporta de la manera esperada.

Objetivo: Verificación y Validación del componente IA

En cumplimiento de los principios relativos al tratamiento[67], se ha de poder demostrar que el componente IA desarrollado trata los datos personales con corrección respetando el principio de exactitud.

Controles:

- El plan de pruebas de verificación incluye revisiones, y en su caso, inspecciones, para detectar y corregir, de manera temprana, defectos en los requisitos, defectos de diseño, especificaciones incorrectas o desviaciones en el desarrollo respecto de los criterios aplicables.

---

[65] Artículo 5.1.b del RGPD – Principios relativos al tratamiento. Limitación de la finalidad Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[66] Artículo 5.2 del RGPD – Principios relativos al tratamiento. Principio de responsabilidad proactiva https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[67] Artículo 5.1.a y 5.1.b del RGPD – Principios relativos al tratamiento. Licitud y limitación de la finalidad https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

- Están contempladas, como parte del plan de pruebas, las pruebas de caja blanca a nivel del diseño de la red o del componente IA[68].
- Están contempladas, como parte del plan de pruebas, las pruebas de caja blanca[69] a nivel de implementación y código.
- Están contempladas, como parte del plan de pruebas, las pruebas de caja negra[70] necesarias para comprobar que la funcionalidad del componente IA está completamente garantizada, su comportamiento es el esperado y que la integridad de la información utilizada se mantiene.
- Están contempladas, como parte del plan de pruebas, las pruebas necesarias para testear la seguridad, con relación a la protección de los derechos y libertades, en su definición holística (física e informática) en el caso de componentes de IA implementados en sistemas robóticos, industria 4.0, o de internet de las cosas[71].
- El plan de pruebas de validación incluye la comprobación de valores límite y casos de prueba extremos que pueden llevar al componente a funcionar de una manera no esperada.
- Existe un proceso de depuración documentado para corregir los errores, carencias o inconsistencias detectadas durante el proceso de verificación y validación.

Objetivo: Rendimiento[72]

En cumplimiento de los principios relativos al tratamiento[73], el componente IA ha de tratar los datos personales respetando el principio de exactitud.

[68] Por ejemplo, en el caso de redes neuronales, estas pruebas incluyen el análisis de la cobertura neuronal, umbral de cobertura, cambio de signo de activación, nivel de cobertura etc. Se pueden encontrar referencias en el esquema de certificación de sistemas de IA del Korean Software Testing Qualification Board: https://imbus.cn/upFile/Uploadfiles/AI%20Testing_Testing%20AI-Based%20System%20Syllabus%20v1.3.pdf

[69] Este tipo de pruebas están fuertemente ligadas al código fuente y resultan clave, desde el punto de vista de la calidad y la seguridad, para hacer una detección temprana de los fallos y vulnerabilidades que presenta el componente en su desarrollo e implementación efectiva. Son necesarias para, en particular, realizar un análisis estático dirigido a determinar la corrección del código generado, la posible manipulación del mismo, la existencia de zonas de código muerto, inalcanzable o redundante, así como la codificación de puertas traseras en las librerías u otros de los elementos empleados en la implementación del componente que pudieran suponer una modificación de las especificaciones funcionales y no funcionales definidas. Suelen seguir un enfoque botón-up, inspeccionando y verificando el comportamiento de cada uno de los componentes de manera individual antes de proceder a su integración dentro del sistema del que forman parte de modo que, una vez validados todos los elementos por separado y realizada la integración, se prueba el comportamiento global del sistema

[70] Entre las pruebas de caja negra que se pueden utilizar se contemplan las de partición de equivalencia, el análisis de los valores límite, las pruebas de tabla de decisión, las pruebas de transición de estado y las de caso de uso.

[71] Véase por ejemplo Mayoral Vilches, V., Olalde Mendia, G., Perez Baskaran, X., Hernández Cordero, A., Usategui San Juan, L., Gil-Uriarte, E., Olalde Saez de Urabain, O. and Alzola Kirschgens, L., 2018. Aztarna, a footprinting tool for robots. *arXiv*, pp.arXiv-1812.

[72] El término "rendimiento" se utiliza en su acepción relativa a eficacia respecto a la protección de datos, antes que respecto otros aspectos o su eficiencia. Dependiendo del tratamiento o el componente se evaluaría la matriz de confusión, la precisión, la sensibilidad, la especificidad o la curva operativa del receptor (ROC) y el área bajo la curva (AUC), basadas en las tasas de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Entre otros, consultar: https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/, https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/, https://imbus.cn/upFile/Uploadfiles/AI%20Testing_Testing%20AI-Based%20System%20Syllabus%20v1.3.pdf

[73] Artículo 5.1.d Principios relativos al tratamiento – Principio de exactitud. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

Controles:

- Están establecidas las métricas o conjunto de métricas agregadas para determinar en el componente su precisión, exactitud, sensibilidad[74] [75] [76] u otro parámetro de rendimiento relativo a la aplicación del principio de exactitud de los datos.
- Son conocidos, y han sido analizados e interpretados, los valores de las tasas de falsos positivos y falsos negativos que arroja el componente IA de cara a determinar la precisión, la especificidad y la sensibilidad del comportamiento del componente.
- Han sido evaluados el nivel y la definición de los parámetros de rendimiento que se requieren para el componente de IA en el marco del tratamiento al que da soporte.
- Han sido comparados los valores de rendimiento entre distintas opciones de componentes IA en el marco de un proceso de elección del componente más adecuado para un tratamiento.
- Están definidas y determinadas las variables de salida prestando especial atención a aquellas que constituyen categorías especiales de datos.
- Han sido adoptadas medidas para garantizar que los datos utilizados son exhaustivos y están actualizados.
- Están determinados los parámetros y sus valores de corte para que el modelo tome en cuenta determinadas variables de cara a obtener resultados que sean significativos.
- Existen procedimientos para detectar si la respuesta del componente IA a los datos de entrada es errónea o supera un umbral de error determinado, o si hay distintos umbrales de error asociados a distintas categorías de interesados en el conjunto de datos.
- Se ha realizado un reajuste de la dimensionalidad[77] del modelo para que exista un equilibrio entre la complejidad y la capacidad de generalización.

Objetivo: Coherencia

En cumplimiento de los principios relativos al tratamiento[78], el componente IA ha de tratar los datos personales respetando el principio de exactitud.

[74] La precisión tiene que ver con lo cerca que está un resultado del valor verdadero, la exactitud con el porcentaje de predicciones correctas, la sensibilidad la fracción de verdaderos positivos (es decir, predicciones correctas de una afirmación correcta) y la especificidad con la fracción de verdaderos negativos (es decir, predicciones incorrectas de una afirmación incorrecta)

[75] Falsos positivos, o la importancia de comprender la información. Cuaderno de cultura científica, 2015. Disponible en: https://culturacientifica.com/2015/10/07/falsos-positivos-o-la-importancia-de-comprender-la-informacion/

[76] Machine Learning: Selección de métricas de clasificación. SitioBigData.com Disponible en: https://sitiobigdata.com/2019/01/19/machine-learning-metrica-clasificacion-parte-3/#

[77] La dimensionalidad se relaciona con la relación entre las variables observadas por el modelo para aprender y el número de muestras utilizadas para el entrenamiento. Si el número de parámetros es muy alto, el modelo se ajustará en exceso a las características de los datos de entrenamiento captando toda la información relevante pero también todo el ruido existente (*overftting* o sobreajuste) por lo que, en las fases de prueba y validación, al realizar predicciones sobre nuevos datos, la precisión será notablemente más baja y no podrá generalizar las reglas para predecir resultados respecto a datos que no ha visto. De ahí la importancia de dividir el conjunto de datos de entrada utilizados en la fase de entrenamiento en dos conjuntos disjuntos (entrenamiento y validación, normalmente en una proporción 80-20) para poder determinar cómo se comporta el modelo con datos que nunca ha visto. http://www.revistaindice.com/numero68/p22.pdf

[78] Artículo 5.1.d Principios relativos al tratamiento – Principio de exactitud. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

Controles:

- Existe un procedimiento para verificar si se producen variaciones significativas en los resultados obtenidos respecto de las salidas esperadas y actuar en caso necesario.
- Se ha establecido un umbral de cara a determinar cuándo un resultado obtenido difiere del esperado ante datos de entrada idénticos o similares (variaciones significativas).
- Se ha determinado si el componente IA se comporta de manera distinta frente a individuos que se diferencian entre sí en características asociadas a categorías especiales de datos o en los valores que toman las variables *proxy*.
- Se han analizado los efectos en los resultados de salida del componente IA ante variaciones de las variables con baja prevalencia en el conjunto de datos de entrenamiento.
- Se han adoptado medidas para garantizar la independencia del componente[79].
- Se ha comprobado que no existe correlación entre los resultados y las variables adicionales asociadas a sujetos[80] que no forman parte de las variables de proceso y que pudieran determinar la existencia de sesgos.

Objetivo: Estabilidad y robustez

En cumplimiento de responsabilidad proactiva[81] el componente IA está sometido a procesos de supervisión continua para adaptarse a las modificaciones producidas en el entorno y detectar las necesidades de reajuste[82] como consecuencia de cambios de contexto internos y externos al tratamiento[83].

Controles:

- Se han identificado los factores, dentro del posible contexto o del contexto real de funcionamiento del componente, cuya variación puede afectar a las propiedades del componente IA y pueden establecer la necesidad de gestionar su reajuste.
- Se ha evaluado el comportamiento del componente IA ante casos de uso o entornos imprevistos.
- Se ha realizado una estimación de tiempos en los que es necesario una reevaluación, reajuste o reinicio del componente para ajustarlo a desviaciones en los datos de entrada o cambios en los criterios de toma de decisiones.
- Está documentado si, por diseño, el componente IA se ha construido siguiendo un enfoque estático o bien un enfoque dinámico o de aprendizaje continuo[84].

---

[79] Un componente IA es independiente si la probabilidad de que genere un resultado no viene determinada por el atributo que define a un grupo específico.

[80] La discriminación que surge debido a la toma de decisiones basadas en características correlacionadas con datos como la raza o el género que se vinculan a grupos protegidos, se conoce como discriminación por proxy. Más información disponible en: https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3572098_code499486.pdf?abstractid=3347959&mirid=1

[81] Artículo 5.2 del RGPD – Principios relativos al tratamiento. Responsabilidad proactiva. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[82] Dicho reajuste puede suponer un reentrenamiento, una nueva evaluación de pesos o restructuración dependiente del tipo de IA que se esté tratando.

[83] Como podrían ser desviaciones en los datos, nuevos riesgos para los derechos y libertades de los interesados o cambios en los criterios de decisión que soportan el modelo

[84] Al construir un componente de IA se puede seguir un enfoque estático, en el que el comportamiento definido durante las fases de entrenamiento y construcción del modelo permanece inalterable conforme a los datos utilizados en el proceso de aprendizaje o, por el contrario, el componente, una vez en producción, utiliza los datos de entrada no sólo para generar un resultado de salida sino también para adaptarse e ir mejorando el modelo.

- En el caso de componentes IA de aprendizaje continuo, se ha evaluado el grado de adaptabilidad a nuevos datos o tipos de datos de entrada y se han definido los procedimientos y mecanismos de supervisión para verificar que las conclusiones extraídas siguen siendo válidas, el componente es capaz de adquirir nuevo conocimiento o y no se está produciendo una pérdida de las asociaciones previamente aprendidas durante el aprendizaje inicial[85].

Objetivo: Trazabilidad

En cumplimiento de los principios de protección de datos[86] y, en particular, del derecho del interesado a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar[87], el comportamiento en el tratamiento del componente IA ha de poder supervisarse a través de mecanismos de trazabilidad, incluidos los medios humanos.

Controles:

- Existe un sistema de control de versiones de todos los elementos del componente IA: conjuntos de datos utilizados, del código del componente, librerías empleadas y de cualquier otro elemento asociado al componente.
- Existe un procedimiento formal, documentado y sujeto a una reevaluación del riesgo en función de aquellos cambios que puedan producirse en la implementación del componente IA a lo largo de su ciclo de vida.
- Existen mecanismos de monitorización y supervisión del componente IA, tales como ficheros de logs y registros de resultados, que permiten evaluar el comportamiento del componente en su interacción con el entorno, medir que estas salidas se ajustan a las respuestas de los procesos de la realidad que modelan y rectificar las posibles inconsistencias que puedan existir entre el comportamiento real esperado y el automatizado.
- Existe un registro de aquellas incidencias y comportamientos anómalos previos detectados y corregidos.
- Los mecanismos de monitorización están disponibles a operadores humanos para su seguimiento y verificación.
- Está documentado un procedimiento para asegurar la intervención humana en la toma de decisiones, tanto de oficio, ante resultados discrepantes en relación con el comportamiento esperado, como de parte, ante solicitud de los interesados afectados por el resultado del componente.
- Están adoptados mecanismos en el marco del tratamiento para que los resultados y las decisiones tomadas puedan llegar a depender, de manera exclusiva, de la responsabilidad de seres humanos.

[85] En este tipo de sistemas hay que evitar lo que se denomina "olvido catastrófico" que consiste en que, según el sistema aprende de nuevos datos de entrada diferentes a los utilizados en la fase de entrenamiento y modifica los parámetros para adaptarse a las nuevas entradas, sobrescribe el conocimiento previamente adquirido en las fases previas de entrenamiento considerándolo inválido.

[86] Artículo 5 del RGPD - Principios relativos al tratamiento. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[87] Artículo 22 del RGPD – Decisiones individuales automatizadas, incluida la elaboración de perfiles. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e2901-1-1

Objetivo: Seguridad

En cumplimiento de los principios relativos al tratamiento[88] y las obligaciones de la protección de datos desde el diseño y por defecto[89] y la seguridad del tratamiento[90], el componente IA ha de tratar los datos personales aplicando, de forma eficaz y efectiva, los principios de protección de datos e integrando las medidas técnicas y organizativas necesarias para garantizar un nivel de seguridad de la información adecuado al riesgo, y en especial, en lo que se refiere a la confidencialidad, integridad, disponibilidad y resiliencia del tratamiento.

Controles:

- Se ha realizado un análisis de los riesgos para los derechos y libertades de las personas a la luz del cual se puedan determinar los requisitos de seguridad y privacidad del componente IA utilizado en el marco del tratamiento.
- Se han definido, en el origen y junto con el resto de los requisitos, aquellos relacionados con la protección de los datos y la seguridad, independientemente de que se trate del diseño de un nuevo componente IA o de la modificación de uno ya existente.
- Se siguen los estándares y buenas prácticas disponibles para el desarrollo y la configuración segura del componente.
- Se han implementado las medidas necesarias para garantizar la protección de los datos tratados, en particular, aquellas orientadas a garantizar la confidencialidad, mediante técnicas de anonimización o seudonimización de los datos y la integridad para proteger la implementación del componente de manipulaciones accidentales o intencionadas.
- Se han implementado medidas para garantizar la resiliencia del componente y su capacidad para resistir ataques[91].
- Se han implementado procedimientos para monitorizar el funcionamiento del componente y detectar, de manera temprana, posibles fugas de datos, accesos no autorizados u otros tipos de brechas de seguridad.
- Los usuarios y operadores del componente disponen de información y conocen sus deberes y responsabilidades en materia de seguridad orientada a la protección de los derechos y libertades de los interesados.

---

[88] Artículo 5.1.f del RGPD – Principios relativos al tratamiento. Integridad y confidencialidad. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e1873-1-1

[89] Artículo 25 del RGPD – Protección de datos desde el diseño y por defecto. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3126-1-1

[90] Artículo 32 del RGPD – Seguridad del tratamiento. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32016R0679&from=ES#d1e3443-1-1

[91] Los ataques de inteligencia artificial consisten en que los terceros no autorizados pueden manipular estos componentes con el fin de alterar su comportamiento para cumplir un objetivo malicioso. El ataque puede afectar a componentes de diseño, librerías, software, hardware y cualquier otro elemento de implementación. Como, cada vez más, este tipo de componentes IA se integran en componentes críticos de la sociedad, estos ataques representan un riesgo potencial que puede tener efectos significativos en la seguridad del país, de ahí la necesidad de identificar sus posibles vulnerabilidades (por ejemplo, limitaciones en los algoritmos o escasos controles en los datos utilizados) así como las fuentes de potenciales amenazas para implementar las medidas de seguridad y buenas prácticas necesarias para garantizar su resiliencia. Comiter M. Attacking Artificial Intelligence - AI's Security Vulnerability and What Policymakers Can Do About It. Belfer Center for Science and International Affairs, 2019. Disponible en: https://www.belfercenter.org/sites/default/files/2019-08/AttackingAI/AttackingAI.pdf

IV. CONCLUSIONES

Como se ha señalado en la introducción, una de las herramientas para “garantizar y demostrar” el cumplimiento del RGPD es la realización de auditorías de los tratamientos y, en un futuro, procesos de supervisión de los códigos de conducta, de acuerdo con lo establecido en el artículo 40 del RGPD y de las certificaciones, de acuerdo con lo establecido en el artículo 42 de la misma norma. Ambos instrumentos requieren disponer de criterios objetivos para la evaluación de la adecuación normativa. Si bien unos criterios generales pueden ser comunes a todos los tratamientos, otros incluirán especificidades por razón, entre otros, de los elementos tecnológicos en los que se apoye el tratamiento. Uno de estos elementos tecnológicos para tener en cuenta es la incorporación al tratamiento de datos personales de componentes basados en inteligencia artificial, bien sea para el desarrollo del componente, bien para su explotación u otras posibles circunstancias.

En este documento se ha realizado una primera aproximación a la determinación de un conjunto de objetivos de control y controles diseñados, desde una perspectiva de protección de datos, para ser incluidos en la auditoría de un tratamiento que incorpora componentes de inteligencia artificial. De cara a la realización de una auditoría, disponer de un listado de objetivos de control y controles, constituye una referencia para poder determinar y contrastar la adecuación del tratamiento objeto de la auditoría, compararlo con otros tratamientos y determinar su evolución. Y, como en toda auditoría, los controles que sean oportunos tener en cuenta, así como la forma de abordar su comprobación, han de ser seleccionados y adaptados por el auditor. La selección de aquellos que sean pertinentes para la auditoría de un tratamiento concreto dependerá de diversos factores: del tipo de tratamiento, de los requisitos del cliente, de la auditoría específica, del objeto y alcance de esta y de los resultados de un análisis de riesgos del tratamiento y del propio proceso de auditoría.

La metodología de auditoría es bien conocida, aunque enfocada a la evaluación de componentes IA puede presentar algunas particularidades que han de tenerse en cuenta, tal y como se señala en el presente documento.

Teniendo en cuenta la evolución que está sufriendo este entorno tecnológico, el presente informe no podría ser sino un documento orientativo y vivo, cuyas futuras versiones tendrán que reflejar la realimentación de su puesta en ejecución.

V. ANEXO I: DEFINICIONES

A los efectos de este documento, resulta útil definir previamente una serie de términos relevantes para facilitar la comprensión de los conceptos desarrollados y que forman parte de los procesos de una auditoría algorítmica de protección de datos.

Anonimización

Siguiendo las especificaciones proporcionadas por el Reglamento (Considerando 26[92] del RGPD), este documento 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 y romper su vínculo con la persona a la que se refieren, de manera que esta no sea identificable a través de ellos.

Aprendizaje de componentes IA

Existen cuatro aproximaciones en los modelos de desarrollo de componentes IA:

- **Aprendizaje supervisado (*supervised Learning*):** un operador actúa como "instructor" del componente, introduciendo datos de entrenamiento en el sistema que contienen los datos de entrada y también los datos de salida "correctos" para esos datos de entrada, es decir, introducen datos etiquetados. El componente debe reproducir este "patrón" en futuras ocasiones, para producir nuevos datos de salida, siguiendo la misma lógica.
- **Aprendizaje no supervisado (*unsupervised Learning*):** A diferencia del aprendizaje supervisado, no existe realimentación por un operador. Los componentes 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.
- **Aprendizaje semi-supervisado (*semi-supervised Learning*):** Estos suponen un compromiso entre los dos anteriores. Contienen algunos datos de entrada etiquetados, aunque la mayoría de ellos no lo están, complementándose con procedimientos automáticos.
- **Aprendizaje por refuerzo (*reinforcement Learning*):** En este caso, el componente está diseñado para observar la interacción del sistema con su entorno y aprovecharla para mejorar el funcionamiento del componente. En el proceso de entrenamiento, 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

[92] Considerando 26: Los principios de la protección de datos deben aplicarse a toda la información relativa a una persona física identificada o identificable. Los datos personales seudonimizados, que cabría atribuir a una persona física mediante la utilización de información adicional, deben considerarse información sobre una persona física identificable. Para determinar si una persona física es identificable, deben tenerse en cuenta todos los medios, como la singularización, que razonablemente pueda utilizar el responsable del tratamiento o cualquier otra persona para identificar directa o indirectamente a la persona física. Para determinar si existe una probabilidad razonable de que se utilicen medios para identificar a una persona física, deben tenerse en cuenta todos los factores objetivos, como los costes y el tiempo necesarios para la identificación, teniendo en cuenta tanto la tecnología disponible en el momento del tratamiento como los avances tecnológicos. Por lo tanto, los principios de protección de datos no deben aplicarse a la información anónima, es decir información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable, o deje de serlo. En consecuencia, el presente Reglamento no afecta al tratamiento de dicha información anónima, inclusive con fines estadísticos o de investigación.

que el sistema toma como “recompensa” y permite determinar cómo de “adecuado” es un determinado comportamiento.

Auditoria

Auditoría es un proceso sistemático, independiente y documentado para obtener evidencias objetivas (registros, declaraciones de hechos o cualquier otra información) y evaluarlas de manera objetiva con el fin de determinar el grado en que se cumplen los criterios de auditoría (conjunto de políticas, procedimientos o requisitos utilizados como referencia)[93].

Auditoría de protección de datos de componentes IA

Aquella parte dentro de una auditoría de protección de datos de un tratamiento cuyo objetivo está limitado a los componentes basados en IA que forman parte de este.

Componente IA

En la primera edición del ISO/IEC TR 29119-11 “*Software and systems engineering - Software testing - Part 11: Testing of AI-based systems*”[94], al igual que en el “*White Paper On Artificial Intelligence - A European approach to excellence and trust*”[95] o en el estudio del Parlamento Europeo “*Artificial Intelligence and Civil Liability*”[96] se utiliza el término “sistema de IA” referido a aquel sistema que incluye uno, o más de uno, componentes basado en IA[97].

Un componente IA es la implementación de un elemento que encapsula las funciones relacionadas con un proceso de inteligencia artificial y que puede incluir los algoritmos, los conjuntos de datos y otros elementos que permiten la ejecución de dicho componente. El componente incluye tanto los aspectos específicos de IA como los que se derivan de su implementación en software y/o hardware, que pueden tener efecto sobre el comportamiento de dicho componente.

El RGPD, tal y como se establece en al artículo 2, aplica a tratamientos. En este caso, el componente IA implementará una fase del tratamiento.

Datos de entrada, datos de salida y datos etiquetados

Se consideran datos de entrada (*input data*) aquellos que se introducen en el componente IA para ser procesados por el mismo.

Los datos de salida (output data), por su parte, son aquellos que resultan del procesamiento algorítmico de los datos de entrada.

---

[93] UNE-EN ISO 19011:2018 - Directrices para la auditoría de los sistemas de gestión. Disponible en: https://www.iso.org/obp/ui#iso:std:iso:19011:ed-3:v1:es

[94] ISO/IEC TR 29119-11:2020 Software and systems engineering — Software testing — Part 11: Guidelines on the testing of AI-based systems. Disponible en: https://www.iso.org/standard/79016.html

[95] Libro Blanco sobre la Inteligencia Artificial - un enfoque europeo orientado a la excelencia y la Confianza [en línea] COM(2020) 65 final, p.30 [Consulta: 29 de octubre del 2020]. Disponible en https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

[96] https://www.europarl.europa.eu/RegData/etudes/STUD/2020/621926/IPOL_STU(2020)621926_EN.pdf

[97] Igualmente, así lo pone de manifiesto CENELEC en “CEN-CENELEC response to the EC White Paper on AI”. Disponible en: https://www.cencenelec.eu/news/policy_opinions/PolicyOpinions/CEN-CLC%20Response%20to%20EC%20White%20Paper%20on%20AI.pdf

Por su parte, los datos etiquetados (*labelled data*) son aquellos datos que se introducen en un componente IA y están vinculados a una determinada información de salida. Las etiquetas en los datos permiten al sistema conocer contenido sobre estos datos[98].

Datos personales

En este documento se utiliza la definición de datos personales[99] proporcionada por el RGPD en el artículo 4.1: "*toda información sobre una persona física identificada o identificable («el interesado»); se considerará persona física identificable 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*".

Desde el punto de vista del ciclo de vida del componente IA (ver definición de ciclo de vida más adelante), podemos encontramos con los siguientes tratamientos de datos personales:

- Cuando se utilizan datos personales en la etapa de desarrollo del componente IA.
- Cuando se utilizan datos personales en las etapas de verificación o validación del componente IA.
- Cuando se incluye el componente IA en un tratamiento de datos personales (etapa de explotación), como podría ser un tratamiento de control de seguridad (que incluye reconocimiento facial) o de atención al ciudadano (que incluye en chatbot).
- En cualquier otra etapa del ciclo de vida que involucre datos personales.

En dichos tratamientos se podrían utilizar conjuntos de datos (*datasets*) y se podrían inferir nuevos datos personales. Todos ellos, directos o indirectos, originales o derivados, son datos personales en tanto y cuanto hagan referencia a un individuo identificado o identificable y por lo tanto objeto de protección de acuerdo con el artículo 1 del RGPD.

Los datos personales se clasifican en identificadores, cuasi-identificadores y categorías especiales de datos.

Los primeros, son aquellos datos que, por sí solos, están asociados de forma unívoca a un sujeto como, por ejemplo, el DNI, el nombre completo, el pasaporte, el número de la seguridad social o cualquier otro identificador que cumpla el mismo propósito.

Los cuasi-identificadores, también llamados identificadores indirectos o seudo-identificadores, son aquellos datos que, sin identificar directamente al individuo, convenientemente agrupados y relacionados con otros conjuntos de datos o fuentes de información, pueden llegar a identificar a una persona y permitir la vinculación o inferencia con datos sensibles. Suelen entrar en la categoría de estos datos la fecha de nacimiento,

[98] En los sistemas de Inteligencia Artificial, el rol del etiquetador es clave para acreditar la validez de los datos y su posterior utilización para ayudar a la máquina en su aprendizaje de cara a lograr que, a medio plazo, los componentes de inteligencia artificial no requieran supervisión.

[99] REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). Diario Oficial de la Unión Europea, de 4 de mayo de 2016. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L:2016:119:FULL&from=DE

el municipio de residencia, el código postal o el género por ser datos muy comunes a un amplio espectro de conjuntos de datos, muchos de ellos de carácter público, en las que puede estar incluido un determinado individuo.

Por último, las categorías especiales de datos son aquellos tipos de datos a los que se les confiere una especial protección de acuerdo con el artículo 9 del RGPD. En concreto son 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 orientaciones sexuales de una persona física."*

Otros datos que, sin entrar en la categoría de especiales, por su naturaleza requieren una especial protección, son los datos personales relativos a condenas e infracciones penales (art. 10 del RGPD), en donde se limita su tratamiento y se establecen garantías especiales como la realización de una evaluación de impacto de protección de datos (art. 35.3 letra b) del RGPD).

Ciclo de vida de un componente IA

El ciclo de vida de un componente de IA es el conjunto de etapas en el que se estructura la evolución de este desde su concepción hasta su retirada.

Las etapas básicas del ciclo de vida de un componente IA podrán ser, en el caso de aprendizaje automático:

- Una etapa de preprocesamiento, en la que se trabaja sobre las bases de datos que servirán para el entrenamiento y la prueba del sistema y que pueden tener datos incompletos, desestructurados y en diferentes formatos, por lo que lo primero de todo será prepararlos para su aprovechamiento. Estos datos se separan generalmente en dos conjuntos: los utilizados para generar el modelo de aprendizaje y los utilizados para su validación.
- Una etapa de preparación del código del componente, que posteriormente se entrena, para generar el modelo algorítmico. La técnica de aprendizaje escogida dependerá de la naturaleza del problema.
- Una etapa de validación sobre el conjunto disjunto de los datos de entrenamiento reservados para este fin.
- Si el modelo se comporta de manera fiable, se procede a las etapas sucesivas de implementación en un modelo comercial, inclusión en un tratamiento, paso a producción, actualización o mantenimiento y finalmente una etapa de retirada.

Es habitual que este modelo de desarrollo cuente con un proceso iterativo de comprobación – reaprendizaje del comportamiento del componente utilizando para ello datos reales que permitan su adaptación y mejora continua.

Discriminación algorítmica

La discriminación algorítmica se refiere al tratamiento desigual que un componente IA da a una persona X con respecto a otra persona Y como consecuencia de un atributo

particular de X. Esta circunstancia no implica, necesariamente, que la discriminación sea negativa o desventajosa[100,101].

Discriminación grupal

Esta forma de discriminación se refiere a aquella discriminación que afecta a una persona a causa de su pertenencia a un grupo socialmente identificable o protegido.

Discriminación estadística

La discriminación estadística se refiere a la discriminación grupal basada en un hecho que es estadísticamente relevante[102].

IA-débil

En función del alcance y el ámbito de aplicación de la inteligencia artificial se diferencian tres categorías de IA: las inteligencias artificiales fuertes, las generales y las débiles. La IA general podría resolver cualquier tarea intelectual resoluble por un ser humano; la IA fuerte o super-inteligencia iría más allá de las capacidades humanas. Por su parte, la IA-débil (*AI-weak*), que es la inteligencia actualmente implementada, se caracteriza por desarrollar soluciones capaces de resolver un problema concreto y acotado. Este documento está orientado a esta última.

Metodología de auditoría

Con independencia del campo doctrinal en el que se aplique, la metodología hace referencia al conjunto de procedimientos racionales, métodos y técnicas que se aplican de manera sistemática durante un estudio o proceso de investigación para alcanzar un determinado objetivo.

En el caso concreto de la auditoría, el objetivo perseguido es determinar el grado de cumplimiento del tratamiento auditado con respecto a los requisitos o criterios de auditoría exigidos, los cuales pueden emanar de las exigencias normativas, las políticas y normas internas del responsable, otros planes definidos en el seno de la organización (plan de calidad, plan de responsabilidad social, …) u otros requisitos especificados por las partes interesadas.

En un proceso de análisis e investigación pueden desarrollarse muchas metodologías, aunque todas ellas caen en dos grandes grupos: la metodología de investigación cualitativa y la cuantitativa.

---

[100] 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.

101 Un ejemplo de discriminación que afecte positivamente a un grupo protegido o vulnerable sería que un componente desarrollado para modelar la asignación de recursos proporcione significativamente más ayudas a personas discapacitadas frente a las asignadas a otros grupos de personas sin discapacidad

[102] Esto puede darse, por ejemplo, en el caso de un componente dedicado a la predicción que utiliza datos sobre probabilidades que proceden del mundo real (y que, por tanto, al reflejar el resultado de decisiones previas, son estadísticamente relevantes), pero cuyo uso da lugar a un tratamiento desventajoso hacia cierto grupo o colectivo social vulnerable. Un ejemplo real de ello es el caso de un componente dedicado a la predicción de reincidencia en la comisión de delitos, que se demostró discriminatorio por su uso de la información relativa a los casos de reincidencia entre las personas de piel negra. 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.

La primera es la que permite acceder a la información a través de la recolección de datos sobre variables, llegando a determinadas conclusiones al comparar estadísticas; la segunda, describe los fenómenos investigados, dejando a un lado la cuantificación de datos y obteniendo la información a través de entrevistas o técnicas no-numéricas, estudiando la relación entre las variables que se obtuvieron a partir de la observación y teniendo en cuenta, sobre todo, el contexto y las situaciones que giran en torno al problema estudiado.

Objetivos de control y controles

De acuerdo con ISO[103], los objetivos de control establecen la declaración de lo que se pretende conseguir mediante la implementación de los diferentes controles, entendidos estos como aquellas medidas por las que se modifica el riesgo. Los controles, por su parte, incluyen procesos, políticas, dispositivos o prácticas, entre otras acciones, para modificar el riesgo. Pueden ser preventivos, detectivos o correctivos en función de cómo actúen de cara a la materialización de la amenaza que conduce al riesgo.

Perfilado

El perfilado (art. 4.4 del RGPD) consiste en aquella forma de tratamiento automatizado de datos personales que permite inferir más información acerca de una persona física, evaluando, analizando o prediciendo aspectos personales.

Un tratamiento que implique la elaboración de perfiles se caracteriza por tres elementos[104]:

- Debe ser una forma automatizada de tratamiento, incluyendo aquellos tratamientos que tienen participación parcialmente humana.
- Debe llevarse a cabo respecto a datos personales.
- Y el objetivo de la elaboración de perfiles debe ser evaluar aspectos personales sobre una persona física.

La elaboración de perfiles y la toma de una decisión sobre una persona física son tratamientos según el Reglamento (Considerandos 24[105] y 72[106] del RGPD) y, en consecuencia, están sometidos a dicha norma.

Riesgo de reidentificación

El análisis de riesgo de reidentificación es un proceso de análisis de datos para encontrar propiedades que puedan aumentar el riesgo de que los sujetos sean identificados. Se pueden utilizar métodos de análisis de riesgos antes de la desidentificación para ayudar a

[103] ISO/IEC 27000 Information technology — Security techniques — Information security management systems — Overview and vocabulary. Disponible en: https://standards.iso.org/ittf/PubliclyAvailableStandards/c073906_ISO_IEC_27000_2018_E.zip

[104] Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679. Grupo del Artículo 29, 2017. Disponible en: https://www.aepd.es/sites/default/files/2019-12/wp251rev01-es.pdf

[105] Considerando 24 del RGPD: "… el potencial uso posterior de técnicas de tratamiento de datos personales que consistan en la elaboración de un perfil de una persona física con el fin, en particular, de adoptar decisiones sobre él o de analizar o predecir sus preferencias personales, comportamientos y actitudes."

[106] Considerando 72: La elaboración de perfiles está sujeta a las normas del presente Reglamento que rigen el tratamiento de datos personales, como los fundamentos jurídicos del tratamiento o los principios de la protección de datos. El Comité Europeo de Protección de Datos establecido por el presente Reglamento (en lo sucesivo, el «Comité») debe tener la posibilidad de formular orientaciones en este contexto.

determinar una estrategia eficaz de desidentificación o después de la desidentificación para vigilar cualquier cambio o valores atípicos.

Sesgo algorítmico

El sesgo[107] algorítmico se produce en aquellos casos en los que un determinado componente IA produce distintos resultados con relación a los sujetos en función de la pertenencia de este a un colectivo concreto (explícito o *ad-hoc*) evidenciando un prejuicio subyacente a dicho colectivo.

Este comportamiento se puede derivar de distintas fuentes: sesgo en los datos de entrenamiento, en la metodología de entrenamiento (p.ej. por una supervisión que incluye el sesgo), por un modelo demasiado simplista (*underfitting*), por una aplicación del componente IA en un tratamiento o un contexto que no es adecuado, etc.

Variables *proxy*

Una variable *proxy* es aquella que se usa en lugar de la variable de interés cuando esa variable de interés no se puede medir directamente[108]. Aunque una variable *proxy* no es una medida directa de la variable deseada, una buena variable *proxy* está fuertemente relacionada con los valores de dicha variable[109].

En definitiva, las variables *proxy*[110] son aquellas que se pueden medir directamente y muestran una correlación lo suficientemente estrecha con aquello que nos interesa pero que no se puede medir directamente, permitiendo así establecer un sistema de evaluación sustitutiva.

[107] Las definiciones de discriminación y sesgo presentadas en esta guía se basan, principalmente, en el trabajo realizado por Barocas y Selbst (2016, "Big data's disparate impact." California Law Review 104: 671.), Baeza-Yates (2018. Bias on the web. Communications of the ACM, 61(6), pp.54-61.), Saldago y Castillo (2018, “Differential status evaluations and racial bias in the Chilean segregated school system.” Sociological Forum, 33, 2, pp. 354-377).), Cowgill y coautores (Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N. and Chaintreau, A., 2020, “Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics.” In Proceedings of the 21st ACM Conference on Economics and Computation, pp. 679-681),Sweeney (2013, "Discrimination in online ad delivery." arXiv preprint arXiv:1301.6822.) .

[108] Proxy variable. Oxford Reference. Disponible https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100351624

[109] Proxy variable. The SAGE Encyclopedia of Social Science Research Methods, 2004. Disponible en: https://methods.sagepub.com/reference/the-sage-encyclopedia-of-social-science-research-methods/n768.xml

[110] Everything is a proxy. Machine Learning: Algorithms in the Real World Coursera [consulta: noviembre 2020]. Disponible en https://www.coursera.org/lecture/machine-learning-applied/everything-is-a-proxy-AFO5D