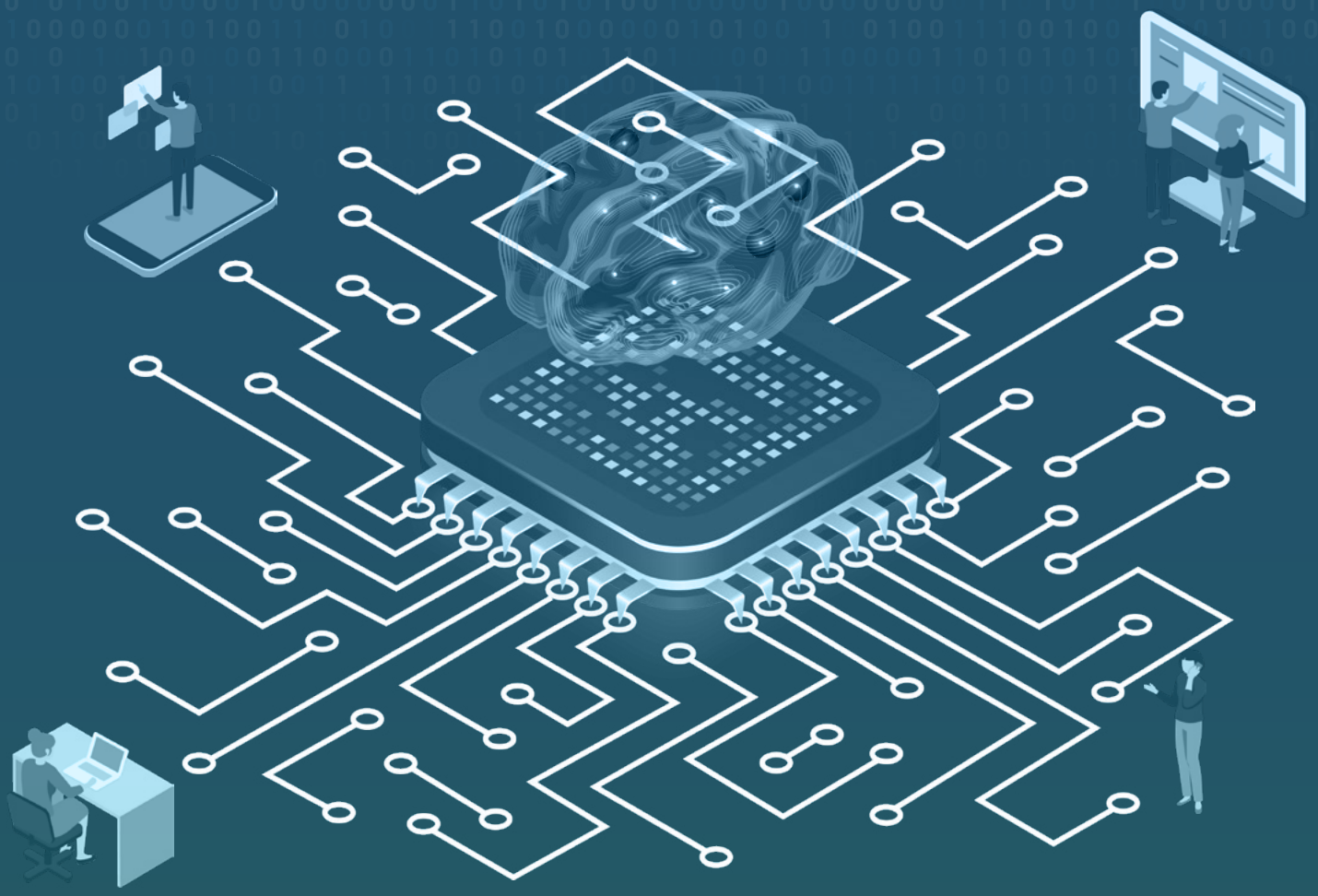


Auditoría algorítmica para sistemas de toma o soporte de decisiones

Matías Aránguiz Villagrán

BID

Auditoría algorítmica para sistemas de toma o soporte de decisiones

Autor: **Matías Aránguiz Villagrán**
Profesor y Subdirector Programa Derecho, Ciencia y Tecnología.
Pontificia Universidad Católica de Chile

Marzo de 2022

Agradecimientos
El autor agradece todos los comentarios de Cristina Pombo y la investigación de Sebastián Dueñas durante la redacción de este documento. Igualmente, agradecimientos especiales a AGESIC, Agencia de Gobierno Electrónico y Sociedad de la Información y del Conocimiento de Uruguay, en especial a Maximiliano Maneiro por su revisión y comentarios, y a Eticas Consulting por las conversaciones iniciales para este documento.

https://www.iadb.org/

Contenido

1. Introducción

La toma de decisiones es una de las habilidades centrales del ser humano. Decidir entre más de una alternativa permite discernir y optar por mejores formas de hacer las cosas. La toma de decisiones es un proceso mediante el cual una persona pondera la información disponible e incorpora su experiencia previa para elegir la opción que, en ese momento, le parece más conveniente.
Daniel Kahneman, Premio Nobel de economía de 2002, distingue en los seres humanos dos formas de pensar que operan en la toma de decisiones: un primer sistema rápido, intuitivo y emocional, y un segundo sistema más lento, deliberativo y lógico. El primero no siempre es eficaz, mientras que el segundo, aunque tarda, permite llegar a conclusiones que incorporan un mayor número de elementos de análisis, un nivel más profundo de reflexión y eficiencia en las decisiones.

Kahneman muestra que la manera de tomar de decisiones de estos dos sistemas es complementaria: la rapidez es esencial en algunas ocasiones, mientras que el análisis complejo y completo es crítico en otras.

Los gobiernos, las empresas, las instituciones y una gran variedad de grupos toman decisiones que inciden la vida de otros (ascensos laborales, beneficios sociales, condenas criminales, etc.).

Por eso la toma de decisiones debe ser un proceso cuidadoso y completo en el cual se debe incorporar toda la información correcta, actualizada y relevante, asegurándose de que todo ello se haga de manera eficiente.

Dado el número de personas afectadas, y el grado de impacto que tienen en las vidas de los afectados[1], los procesos de toma de decisiones gubernamentales deben ser realizados con especial cuidado e incorporar criterios de participación democrática y rendición de cuentas.

1 Por cuestiones estrictamente de estilo, en este documento se usa el género masculino no marcado inclusivo, independientemente del sexo de las personas.

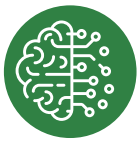
1.1 ¿Qué es un sistema automatizado de toma o soporte de decisiones?

Un sistema automatizado de toma o soporte de decisiones (ADS por las siglas en inglés de Automated Decision Support) es un sistema computacional que puede, para un determinado conjunto de objetivos definidos por seres humanos, hacer predicciones y recomendaciones o tomar decisiones que inciden en entornos reales o virtuales. Estos sistemas están diseñados para operar con diversos grados de autonomía[2].

Durante los últimos años, los ADS han crecido exponencialmente en número y ámbitos de aplicación. En la actualidad, cada vez interactuamos con una mayor cantidad de ADS, a menudo sin que nos percatemos de ello. Sin embargo, la falta de conciencia acerca de su uso no reduce los riesgos sociales, en caso de que estos sistemas estén mal diseñados o se hayan creado sin tomar las precauciones necesarias.

Si los ADS se usan con grupos o comunidades vulnerables como niños, personas con discapacidad y poblaciones históricamente desfavorecidas o en riesgo de exclusión,[3] será necesario tener incluso mayores previsiones en el momento de su implementación.

En esta guía, los ADS se definen en dos grupos según su grado de autonomía:



ADS en los que la información generada por los **modelos de aprendizaje automático** se utiliza como insumo para **la toma de decisiones por parte de una persona**.

ADS en los que las decisiones finales y sus acciones derivadas se toman **sin intervención humana directa**[4].

Esta guía no pretende ser únicamente un instrumento práctico de identificación y mitigación de riesgos o peligros que quizás no sean patentes a simple vista. Se trata igualmente de que sirva como instrumento que ayude a tomar conciencia sobre las implicaciones y consecuencias que conlleva la puesta en marcha de sistemas automatizados en la toma o soporte de las decisiones que afectan las vidas de las personas.

1.2 ¿Para qué sirve una auditoría?

Por lo general, cualquier sistema puede presentar fallas o riesgos que no se detectan a primera vista o cuya relevancia se descuida debido a la frecuencia con que se realizan ciertos procesos. Mientras más complejos sean los sistemas, existen mayores probabilidades de que se presenten errores. Al mismo tiempo, la complejidad de los sistemas permite a menudo una mayor adaptabilidad a la realidad sobre la cual estos hacen predicciones.

Según la norma ISO 19.011 sobre "Directrices para la auditoría de sistemas de gestión", una auditoría debe ser un proceso sistemático, independiente y documentado con el cual se busca recolectar evidencias y evaluarlas para determinar el grado en que se cumplen ciertos criterios previamente determinados[5].

Una auditoría debe incorporar los objetivos de la entidad, la protección de los intereses y necesidades de beneficiarios, colaboradores y otras posibles partes interesadas, así como los requisitos de seguridad y privacidad de la información[6]. Es así como existen auditorías de distinta naturaleza: las hay de tipo contable, legal y de procesos e informática, entre otras. La utilidad de las auditorías radica en que permiten que se haga una evaluación objetiva de los posibles riesgos, se los cuantifique y se priorice su mitigación.

2 OECD, Recommendation of the Council on Artificial Intelligence. Disponible en https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

4 BID. Guía de aplicación Autoevaluación ética de IA para Actores del Ecosistema Emprendedor. Disponible en: https://publications.iadb.org/publications/spanish/document/Autoevaluacion-etica-de-IA-para-actores-del-ecosistema-emprendedor-Guia-de-aplicacion.pdf

4 BID: IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial. Disponible en https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

5 ISO 19011-2018, disponible en https://www.iso.org/standard/70017.html

6 ISO 19011-2018, disponible en https://www.iso.org/standard/70017.html

Si bien las auditorías se han convertido en un componente fundamental en el campo cada vez más extenso de la gobernanza algorítmica[7], por sí solas no son suficientes para mitigar los impactos de la implementación y ejecución de un sistema; en esencia constituyen un proceso mediante el cual se determina el cumplimiento de ciertos estándares. Con todo, las auditorías sí desempeñan un papel indispensable en materia de evaluación de impacto y de consecución y disponibilidad de información, tanto en el ámbito de la entidad misma, como en el de los organismos reguladores, así como para los afectados potenciales y para la sociedad en su conjunto.

1.3 ¿En qué consiste una auditoría algorítmica?

Una auditoría algorítmica es un estudio que busca evaluar un ADS y su proceso de desarrollo, incluyendo el diseño y los datos utilizados para entrenar el sistema. Asimismo se evalúan los impactos en materia de precisión, justicia algorítmica,[8] sesgos, discriminación, privacidad y seguridad, entre otros[9].

Las auditorías algorítmicas se pueden realizar a manera de medición frente a ciertos estándares (auditorías de rendimiento), o bien como un análisis de cumplimiento de normas particulares (auditorías de cumplimiento) con el propósito de producir recomendaciones en materia de métricas específicas[10].

1.4 ¿Para quién es esta guía?

Esta guía está dirigida a los responsables por la formulación de políticas de América Latina y el Caribe y/o a los encargados de liderar proyectos de ADS que tengan a su cargo la mitigación de los impactos producidos por su uso. La idea es que este documento sirva como guía para supervisar estos desarrollos desde las etapas previas al diseño, pasando por la implementación, hasta los posibles ajustes y actualizaciones necesarios para el adecuado uso del modelo de inteligencia artificial. Se trata de apoyar al lector orientándolo acerca de la necesidad de hacer auditorías a los sistemas de inteligencia artificial e indicándole cuáles son los elementos a considerar durante su realización.

1.5 ¿Cómo usar esta guía?

Con este documento se busca introducir al lector al tema por medio de preguntas estructuradas para tomar la decisión de la implementación de una auditoría y del proceso que esta conlleva. La guía se debe usar como acompañamiento durante el ciclo de vida del sistema:[11] desde su conceptualización y diseño, pasando por su uso, hasta la necesaria rendición de cuentas. Además se incluyen referencias para quienes estén interesados en conocer ciertos tópicos en profundidad, lo cual permitirá hacer énfasis en aquellos que resulten particularmente relevantes según el tipo de entidad, el origen de los datos y/o el modelo utilizado, entre otros factores.

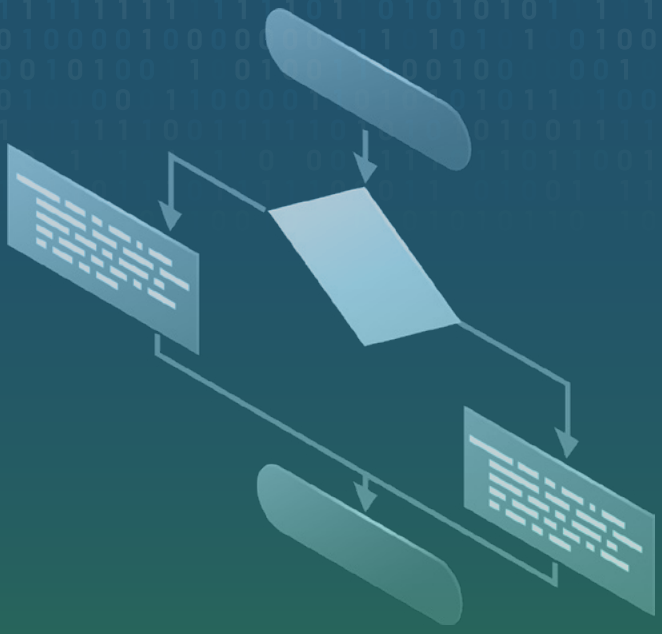
7 Ada Lovelace Institute. *Examining the Black Box: Tools for assessing algorithmic systems*. disponible en https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

8 En este contexto, por justicia algorítmica se entiende la característica de un algoritmo que permite que, al ser aplicado, no produzca daño o discrimine a una persona o grupo.

9 Algorithmic Accountability Act of 2019

10 INTOSAI. *Performance Audit Principles*, disponible en https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_300/ISSAI_300_en_2019.pdf

11 Uso responsable de IA para política pública: manual de formulación de proyectos, disponible en https://publications.iadb.org/es/uso-responsable-de-ia-para-politica-publica-manual-de-formulacion-de-proyectos



2. La auditoría algorítmica

2.1 ¿Por qué realizar una auditoría algorítmica?

Con la adopción generalizada de los ADS en los sectores público y privado, son cada vez más las dimensiones de la vida de la gente que se encuentran bajo su influjo. Desde el tiempo de espera de los usuarios del transporte público, hasta la asignación correcta de servicios estatales, en todo ello se busca que el beneficio sea óptimo.

La implementación de sistemas automatizados conlleva desafíos que a menudo no se abordan con la suficiente profundidad. Entre ellos se encuentran, por ejemplo, las vulneraciones a derechos fundamentales por el uso de datos personales, la discriminación no deseada por parte de entidades e incluso la toma de decisiones difíciles o hasta imposibles de justificar.

A la luz de lo anterior, resulta imperativo tomar medidas de control y revisión tanto internas como externas, especialmente en el sector público. Aquí las auditorías algorítmicas resultan de suma

utilidad, dado que se trata de procesos prácticos y eficaces realizados por terceros cuya finalidad es garantizar que las decisiones se tomen de manera correcta a la luz de consideraciones éticas y técnicas, al tiempo que se respetan los derechos de los ciudadanos.

Si bien se trata de una materia cuya normativa sigue en discusión y desarrollo en diversos países, las distintas políticas nacionales relativas a la inteligencia artificial, así como las diversas directrices internacionales, ponen de manifiesto la necesidad de contar con mecanismos de control adecuados.

En los organismos públicos, la realización de estas auditorías permite verificar que se cumplan los siguientes propósitos[12]:

Respetar el derecho de los ciudadanos a saber con qué sistemas se encuentran interactuando en sus vidas, enumerando y describiendo públicamente los ADS que afectan significativamente a personas y comunidades.

Fortalecer la capacidad interna de los organismos públicos de evaluar los sistemas que construyen o adquieren, y facilitar que obtengan una mayor experiencia en estas tareas. Esto para que logren anticipar los problemas que puedan surgir de situaciones indeseadas, entre ellas la asignación incorrecta de beneficios o las violaciones al debido proceso, por mencionar solo dos.

Garantizar una mayor responsabilidad en el uso de los ADS diseñando un método útil y continuo para que terceros revisen y evalúen estos sistemas, de modo que sea posible identificar problemas y resolverlos o mitigarlos.

Garantizar que la ciudadanía tenga una oportunidad de réplica y, de ser necesario, de impugnar el uso de un determinado sistema o los lineamientos empleados para su desarrollo por parte de un organismo público.

Si, por el contrario, no se realiza este tipo de auditorías, el uso incorrecto de los ADS podría conllevar desde un aprovechamiento no óptimo de los recursos, hasta el desencadenamiento de casos de vulneración de derechos fundamentales de diversos sectores de la población. Los riesgos y daños potenciales son variados y a menudo difíciles de anticipar. Los hay fundamentalmente de dos tipos: riesgos de inclusión (asignación de recursos o beneficios a quienes no corresponde) y de exclusión (privación de recursos o beneficios a personas que sí los necesitan).

La implementación de ADS sin controles y/o auditorías también puede causar un daño reputacional para quienes implementan el sistema, fracturando así la confianza que la sociedad deposita en su actuar diligente y correcto. Esto también puede dar lugar a una desconfianza generalizada en la tecnología, haciendo que la población se muestra cada vez más reacia al uso de ADS en el sector público.

Cabe señalar igualmente la existencia de diversos riesgos propios del desarrollo de las herramientas de inteligencia artificial.

12 IA Now Institute. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.* disponible en https://ainowinstitute.org/aiareport2018.pdf

Aquí figuran, por ejemplo, el énfasis exagerado en la optimización de métricas de rendimiento específicas, en detrimento de las dimensiones de transparencia y equidad. Otro riesgo patente emana de la falta de recursos necesarios para desarrollar los modelos internamente en los organismos que los requieren, y que con frecuencia optan por comprar herramientas que, si bien fueron diseñadas por terceros para usos diversos, terminan siendo adaptadas para el fin particular del adquirente. A ese, se le puede agregar el riesgo que los datos en los cuales se basan el sistema no sean representativos de igual forma para todos los casos, creando un sistema donde la desigualdad es lo correcto. Esto puede crear dificultades en la adecuación del modelo, dados sus requisitos operativos y las exigencias regulatorias para su funcionamiento.

Aun así, la realización de auditorías algorítmicas permite que las entidades satisfagan las exigencias que en materia de eficiencia y eficacia deben cumplir tanto los organismos públicos como los privados, ya sea por la normatividad existente o por las exigencias de la ciudadanía en materia de transparencia y eficiencia.

2.2 Requisitos para realizar una auditoría algrítmica

Los ADS pueden ser desarrollados internamente por el servicio o entidad gubernamental o por terceros. En este último caso su desarrollo se formaliza mediante contrato de adquisición de productos o servicios, licitación o compra directa. En el momento en que se toma la decisión de adquirir un ADS, es clave que la institución cuente con un encargado de proyecto que pueda administrar el contrato y que tenga algún nivel de conocimiento técnico para no perder el control del proceso de desarrollo e implementación.

Al adquirir el servicio de un tercero, será necesario incorporar en el contrato de licitación o de compra directa cláusulas que permitan auditar el sistema. Es conveniente no limitar el número de auditorías ni sus condiciones, a fin de no impedir su realización cuando se requiera. Al mismo tiempo, en la licitación debe dejarse claro que la auditoría puede hacerse directamente o por un tercero a nombre del organismo público.

Algunas empresas pueden mostrarse reacias a ser auditadas, especialmente si existe la posibilidad de que la información se haga pública. Por ello se debe exigir que haya acceso al código fuente de los ADS, de modo que se los pueda auditar ex post. Cabe señalar que las normas sobre cumplimiento regulatorio de los proveedores nunca se pueden esgrimir como limitante para que se conduzcan auditorías.

Además, en las especificaciones de la licitación debe agregarse como requisito para la prestación del servicio que los ADS deberán acompañarse de toda la documentación técnica relativa a su desarrollo. La documentación comprende desde manuales de uso y políticas hasta descripciones técnicas del proceso de entrenamiento, diseño e implementación. Todo ello es fundamental para poder llevar un registro que permita a los auditores revisar los ADS.

2.3 ¿Cuándo se debe realizar una auditoría algorítmica?

A diferencia de otro tipo de evaluaciones, las auditorías algorítmicas se realizan con posterioridad a la implementación del sistema, cuando este se encuentra en operación. De esa forma se puede contrastar el diseño y desarrollo de un ADS con los efectos de su implementación, particularmente cuando ya han existido casos en que se haya producido algún riesgo o daño.

Determinar el momento exacto para realizar una auditoría algorítmica no es un ejercicio trivial. Esto por cuanto, por lo general, los efectos de un ADS en la población se hacen evidentes para el líder del proyecto solo después de que se han causado los daños. Si no han ocurrido, se recomienda enfáticamente realizar las auditorías al finalizar el periodo piloto del proyecto, el cual tiene lugar cuando se hace una implementación controlada en una muestra del universo total. Por ejemplo, si se va a implementar un ADS que ayuda a calificar el riesgo socioeconómico de las familias para asignarles beneficios sociales, el proyecto piloto debería iniciarse en una localidad pequeña antes que en una región o en todo el país.

Al finalizar el proyecto piloto, es decir, al haber hecho pruebas de implementación en el periodo determinado y sobre una muestra definida, se

recomienda evaluarlo. Se pueden considerar escenarios previos a la implementación del piloto, por ejemplo, en una simulación donde se puedan detectar posibles errores en base a diferentes escenarios resultado. Si la evaluación preliminar interna a cargo del equipo que desarrolló el ADS revela la existencia de complicaciones, se recomienda proceder a la auditoría.

Criticidad de los sistemas

La criticidad se refiere a la **importancia** y el **riesgo** de los ADS en su diseño e implementación. Por **importancia** se entiende la función que cumple un ADS, ya sea dentro de una cadena de procesos durante la cual se alimenta a otros sistemas con la información generada, o por el papel que desempeña en una labor determinada. Entre los ejemplos de sistemas críticos figuran aquellos que directamente asignan derechos, ayudas o subsidios, o los que operan en áreas que por su naturaleza son sensibles como por ejemplo la defensa nacional, la salud o el sistema de prisiones.

Por **riesgo** se entiende la posibilidad de que, durante su utilización, un ADS pueda cometer errores que produzcan un daño a la población involucrada. Tal sería el caso de aquellos ADS que toman decisiones sobre la libertad de un imputado, o sobre la asignación de recursos con fines sociales o sobre la respuesta a situaciones de conflicto. Al analizar el daño deben tomarse en consideración por lo menos tres elementos: (i) la **probabilidad de que el daño ocurra**, que por lo general se mide determinando cuán preciso es el sistema para cumplir con su tarea o, en el extremo opuesto, estableciendo cuántas veces se equivoca; (ii) la **profundidad del impacto**, es decir, las consecuencias leves o graves que pueda tener el error (Los errores más graves serían aquellos que inflijan daños a la vida, libertad o propiedad de un individuo o un grupo social, o aquellos que los priven de un servicio o ayuda pública esenciales para su supervivencia), y (iii) la **distribución del error**, es decir, cuando el error que comete el ADS afecta más a un subgrupo de la población que a otros, como ha sucedido con segmentos de bajos ingresos, inmigrantes y minorías raciales. Un ejemplo que ocurre constantemente se da con

sistemas de reconocimiento facial que tiene un funcionamiento óptimo con personas de piel más clara y tiende a cometer más errores con personas de pieles más oscuras. La discriminación sobre dichos grupos causan injusticias en asignaciones o clasificaciones y producen un malestar importante en los individuos del subgrupo discriminado.

Al analizar la información de ambos grupos de elementos, a saber, importancia y riesgo, es posible determinar la criticidad de un sistema. Mientras más crítico sea, mayores serán las precauciones que se deben tomar. Algunos sistemas pueden generar un riesgo tal que su implementación no se justifique, como ocurre con aquellos cuyo nivel de precisión es muy bajo y afectan directamente el bienestar de un grupo de la población. En estos casos, en la deliberación sobre la conveniencia de su implementación se debe sopesar la opción de no hacerlo.

En aquellos casos en que la implementación sea fundamental y ello signifique un riesgo soportable, se deberán tomar medidas de mitigación y control que logren reducir suficientemente los errores y sus efectos. Aquí existen distintas opciones: desde revisiones constantes de los resultados del sistema, hasta la intervención humana en las decisiones sobre cierto subgrupo social o sobre todos los intervinientes, pasando por la transparencia algorítmica, entre otras.

La realización periódica de estas auditorías también es altamente recomendable, sobre todo en contextos sociales en flujo constante y/o en sistemas cuyo funcionamiento pueda evolucionar conforme a una mayor cantidad y variedad de datos utilizados, entre otros. La periodicidad de estas auditorías también debe determinarse considerando el riesgo de error que presente el sistema. En caso de percibir errores o efectos negativos debido a su uso, la realización de una auditoría periódica se convierte en un imperativo.

Cabe notar que también habrá casos en que las auditorías no logren cumplir su propósito, por haber sido realizadas en momentos que no resultan óptimos, a saber:[13]

- **Auditorías prematuras:** En estos casos, la auditoría se ha realizado en una etapa muy temprana, antes de que aspectos importantes del sistema hayan sido completamente implementados o de que sea posible evaluar correctamente los daños posibles.
- **Auditorías tardías:** Dado que las auditorías son un ejercicio ex post, este es útil para remediar situaciones futuras, pero no para abordar los daños acaecidos con anterioridad. Por ello, a menudo se detectan situaciones en las cuales se han causado daños por un periodo considerable, sin que estos fueran evaluados y/o mitigados cuando correspondía.
- **Auditorías esporádicas:** Las auditorías son un mecanismo en constante desarrollo que evoluciona y madura según los avances de la tecnología y de la sociedad. Por ello puede ocurrir que los impactos potenciales anticipados en una auditoría que se realice durante los primeros meses desde la implementación del sistema no sean los mismos que aquellos que se presentan en etapas posteriores. Se insiste entonces en la recomendación de realizar periódicamente estos procedimientos, con el fin de mantener actualizado el mecanismo de la auditoría y la información que suministra la entidad involucrada.

2.4 ¿Quién debe realizar una auditoría algorítmica?

Para evaluar la efectividad y las consecuencias potenciales del sistema, estas auditorías pueden ser realizadas tanto externa como internamente. Según quién la realice, se pueden clasificar en tres tipos[14]:

- **Auditorías a cargo de terceros:** Son aquellas realizadas por agentes externos a la entidad auditada, quienes evalúan el comportamiento de un sistema con base en únicamente en sus resultados.
- **Auditorías a cargo de segundos:** Son aquellas realizadas por un proveedor, cliente o contratista de la institución auditada a los que se otorga acceso al servidor (backend) del sistema y evalúan su comportamiento considerando tanto la parte técnica como los resultados.
- **Auditorías internas:** Son aquellas realizadas por un miembro o equipo de la entidad involucrada, con el fin de evaluar las preocupaciones propias de aquella. Por lo general, tales preocupaciones se originan en los desafíos comunes que entraña el desarrollo responsable de sistemas de inteligencia artificial como son la transparencia y la equidad. Este tipo de auditorías busca alcanzar metas relativas al sistema en sí mismo, considerando sus propios criterios de éxito.

Cabe señalar que, en las tres categorías anteriores, e independientemente de si el auditor es interno o externo, un requisito indispensable y común es que quien quiera que esté a cargo de realizar la auditoría no debe haber estado involucrado en el desarrollo del sistema.

Es posible que se realice más de una auditoría, en más de una de las modalidades indicadas. En dichos casos será relevante no duplicar esfuerzos cuando no se justifique y que las auditorías sean complementarias entre sí.

Dentro de las competencias que debe poseer el equipo auditor se encuentran: (i) conocimientos

13 Assembling Accountability: Algorithmic Impact Assessment for the Public Interest, disponible en https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

14 Assembling Accountability: Algorithmic Impact Assessment for the Public Interest, disponible en https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

técnicos, (ii) conocimientos sobre el área específica donde se implementa el ADS, y (iii) un conocimiento sólido acerca de los principios éticos que deben incorporarse en los sistemas autónomos.

Que el equipo auditor tenga **conocimientos técnicos** significa que domina los lenguajes de programación y las metodologías específicas usadas en los ADS, y que puede validar la selección y el trabajo de datos. Por ejemplo, si alguien que se hace llamar auditor no es especialista en la tecnología de IA utilizada en el ADS de la entidad, poco podrá decir sobre el cumplimiento auditable de dicho sistema.

El **conocimiento sobre el área específica donde se implementa el ADS** es fundamental para que el equipo auditor pueda evaluar y hacer recomendaciones sobre mejoras en el sistema. Por ejemplo, si se trata de un sistema que evalúa la peligrosidad de individuos que tienen procesos penales (véase el caso de COMPAS en la sección 5), y para ello se toma en cuenta el color de la piel del individuo imputado y/o el barrio de donde proviene --lo cual es de por sí contrario a los derechos humanos--, los auditores deberán ser capaces de advertir el sesgo si se quieren cumplir con el propósito de mejorar el sistema.

El equipo auditor debe tener un **conocimiento sólido de los principios éticos que se deben incorporar en los sistemas autónomos** para poder evaluarlos durante su revisión. Hoy existen distintos marcos éticos que así lo permiten, como se verá más adelante. Igualmente es necesario adoptar principios en materia de protección de datos personales y herramientas que contribuyan a evitar o a reducir los casos de discriminación. El equipo auditor debe tener la capacidad de determinar si se cumplen o no las normas y principios éticos que deben regir el funcionamiento de los ADS para proteger la dignidad de las personas, y no enfocarse solamente en la eficiencia o en las fallas del sistema.

Desde el punto de vista del diseño de la auditoría, es necesario establecer canales de comunicación entre el equipo responsable de esta y el de la institución auditada, prestando especial atención a los cargos y funciones que tienen que ver con el desarrollo del sistema y que se describen en el numeral 2.10.

2.5 ¿Qué información se comparte, y cómo, en un proceso de auditoría?

Para poder auditar un sistema, primero debe existir un registro de la documentación detallada acerca de los procesos de entrenamiento, de la realización y validación de las pruebas, y de su implementación. Cuanto más detallada sea la documentación, más fácil será el trabajo de los auditores. Sin embargo, el costo y los esfuerzos invertidos en documentar el ciclo de vida del sistema autónomo dependerán del nivel de criticidad del proceso.

La documentación detallada permite que los auditores revisen la historia del desarrollo del algoritmo, incluyendo los fines para los que fue creado, el equipo que trabajó en él, las pruebas realizadas y las modificaciones que ha sufrido. De esta manera se puede comparar el sistema en las diferentes etapas de su vida, lo cual es sumamente útil para determinar la instancia exacta en la que se podría producir una anormalidad.

2.6 ¿Qué debe incluir la documentación?

La realización de una auditoría algorítmica conlleva varios procesos de evaluación. Estos abarcan desde el **modelo de gobernanza de la entidad** cuyo sistema va a ser auditado (organigrama y funciones del equipo involucrado en su desarrollo, planes estratégicos para su uso e implementación, partes interesadas y afectadas por ello, entre otros elementos), las **bases de datos** utilizadas (método de recolección de datos, y calidad, pertinencia, manejo y manipulación de los mismos, entre otros elementos), hasta llegar al **modelo computacional** (algoritmos utilizados, sensibilidad y especificidad del sistema e impacto y distribución del error, entre otros elementos). Por lo tanto, la documentación entregada por la empresa auditada deberá permitir entender el modelo de gobernanza del sistema, así como construir un perfil adecuados de datos[15] y del modelo algorítmico propiamente tal[16].

15 BID. IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial, pág. 55. Disponible en https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

16 BID. IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial, pág. 57. Disponible en https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

Para construir el perfil de los datos utilizados en el modelo será necesario tener la información acerca de su origen, recolección, gobernanza y estructura, además de una evaluación de su calidad. Para elaborar el perfil del modelo será necesario contar con la información acerca de su conceptualización y diseño; sobre la fuente y el manejo de los datos; sobre su desarrollo, uso y monitoreo, y sobre la correspondiente rendición de cuentas. El documento "Uso responsable de la IA para las políticas públicas: manual de ciencia de datos"[17] ofrece una aproximación pormenorizada de los ítems contenidos en cada uno de ellos.

2.7 ¿En qué condiciones se le deben entregar los datos al auditor?

Dependiendo del nivel de criticidad del sistema, será importante definir cuáles son las condiciones para el traspaso de la documentación. Para ello deberán definirse explícitamente los usos permitidos y prohibidos tanto de los algoritmos como de los datos. Aquí será de gran utilidad contar con un Acuerdo de Transferencia de Datos (*Data Sharing Agreement*). Este documento permite establecer la responsabilidad y las funciones de cada una de las partes; clarificar el propósito de la transferencia de datos; detallar lo que ocurre con estos en cada etapa; y establecer estándares de uso, seguridad y privacidad.[18] De esta manera, tanto el auditor como el auditado contarán con un documento que esclarezca las responsabilidades de cada parte en materia de datos. Lo anterior es particularmente relevante, por ejemplo en casos en que la confidencialidad sea crítica, como aquellos donde se manejan datos personales, información de seguridad nacional u orden público o información comercialmente sensible.

Los datos de entrenamiento podrán ser entregados a los auditores para que ellos puedan reproducir el proceso y evaluar si existe una mejor forma de trabajar con dicha información. Para realizar este procedimiento será fundamental que se cumpla la normativa en materia de protección de datos personales, en caso de que aplique. Por ejemplo, sería conveniente anonimizar la base de datos con el fin de proteger completamente tal información. Asimismo, en el acuerdo de transferencia se debe especificar que estos datos tendrán la auditoría como único propósito y que no se usarán para fines distintos a ella.

2.8 ¿Quién debe tener acceso a los resultados de la auditoría?

La regla general de la administración pública es el **principio de transparencia**, según el cual los actos, resoluciones, procedimientos y documentos de la administración del Estado deben ser públicos. Este principio permite que haya rendición de cuentas del Estado para con la sociedad civil, representada directamente por organizaciones sociales o comunitarias, universidades y centros de estudios.

En el caso de los resultados de una auditoría algorítmica, es importante determinar quiénes son los terceros que van a tener acceso a la información y a la evaluación preparada por el o los auditores. El informe de auditoría contendrá un análisis de la eficacia de los algoritmos, pero también puede mostrar su funcionamiento, los tipos de datos que se usan y las posibles vulnerabilidades. De allí que sea crítico analizar y determinar el grado de divulgación de los informes.

Para ello deberá identificarse, primero, qué información del proceso automatizado es sensible y cuál puede ser libremente conocida por terceros. Paso seguido se debe identificar cuáles de los elementos que van a analizar los auditores pueden constituir un riesgo para la continuidad operacional del sistema y para la protección de los beneficiarios. Por último, no se debe pasar por alto que la retroalimentación por parte de la sociedad civil permite mejorar los procedimientos en direcciones no necesariamente previstas y revisar democráticamente los procesos que tienen efectos sobre las vidas de las personas. Es recomendable que la retroalimentación de la ciudadanía y sociedad civil pueda hacerse por medios directos, donde exista, al menos, un correo electrónico donde se puedan recibir los reclamos o sugerencias. La transparencia no solo se satisface por medio de la publicación de información, sino que con mecanismos de

17 Ibid.

18 UK Information Commissioner's Office. Data sharing code of practice. Disponible en: https://ico.org.uk/media/for-organisations/guide-to-data-protection/ico-codes-of-practice/data-sharing-a-code-of-practice-1-0.pdf

participación en que las personas puedan hacer valer sus inquietudes de forma directa a la autoridad.

La sensibilidad de la información contenida se determinará dependiendo de si su conocimiento puede producir o no un daño a los beneficiarios, incidir adversamente en la continuidad operacional del sistema o afectar la eficacia del servicio. Por el contrario, no deberá considerarse sensible y/o confidencial aquella información que contenga sesgos sobre ciertos grupos o que evidencie discriminación. Esta deberá hacerse pública para que los beneficiarios puedan defenderse y proteger sus intereses en caso de que se hayan registrado errores.

Dependiendo de la necesidad de mantener la confidencialidad, las auditorías se pueden clasificar según sus grados de transparencia y divulgación. Un caso de **máxima confidencialidad** (menor divulgación) sería aquel en que los resultados de la auditoría sean conocidos únicamente por el organismo que está implementando el sistema. Un caso de **mediana confidencialidad** sería aquel en que la misma información pueda ser compartida con organismos superiores jerárquicos o evaluadores. Un caso de **menor confidencialidad** se daría cuando la información utilizada pueda ser compartida con organismos públicos pares que puedan beneficiarse de ella para mejorar sus propios procesos.

El nivel **mínimo de confidencialidad** es aquel en el cual se comparte la información con terceros ajenos a la administración pública como pueden ser las instituciones internacionales, las universidades o los centros de estudios. Aquí la integridad de los datos puede asegurarse verificando que la información compartida se rija por acuerdos de confidencialidad (también conocidos como NDA por las siglas en inglés de Non-disclosure Agreements). Lo anterior garantizaría que la información no se divulgue o se use para fines distintos a los de la auditoría.

Por último, la información será de divulgación amplia cuando su conocimiento no constituya un peligro para la continuidad operacional del sistema, para las personas y/o para la eficiencia del servicio. También es posible que se haga pública la auditoría cuando los riesgos identificados ya hayan sido neutralizados y se hayan saneado las vulnerabilidades.

2.9 Consideraciones para realizar una auditoría algorítmica

Dentro de los supuestos que subyacen a la realización de una auditoría algorítmica figuran las siguientes:

(i) Quien audita –trátese de un ente interno o externo a la institución-- es independiente y ajeno al desarrollo e implementación del sistema.

(ii) Quien desarrolla e implementa el sistema debe ser capaz de suministrar información adecuada sobre este a quien realiza la auditoría.

(iii) Quien audita debe poder entender correctamente el sistema con base en la información proporcionada, la documentación pertinente y los efectos que pueda percibir respecto a los impactos del sistema.

(iv) El comportamiento del sistema durante su uso y monitoreo es consistente con su comportamiento en el momento de ser auditado[19]. Es de suma relevancia tener esto en cuenta, dado que su funcionamiento puede variar dependiendo del contexto o de los datos con los que se lo alimente. Esto último es lo que justifica la necesidad de realizar estas auditorías de manera periódica para dar cuenta de posibles cambios de escenario, de la inclusión de nuevas funciones o de la supresión de otras.

(v) Siempre que sea posible, la auditoría se debe realizar en términos binarios, lo cual significa que las evaluaciones deben expedirse en un formato que no permita matices (por ejemplo, cumple/no cumple). La razón para que así sea es que una graduación de la evaluación puede llevar a áreas grises que socaven la claridad y confianza requerida en la auditoría.

19 Ada Lovelace Institute. Algorithmic Accountability for the Public Sector. Disponible en: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

2.10 Determinación de daños

Durante la auditoría se revelará —y se podrá medir— el daño que ha generado un sistema debido a un funcionamiento defectuoso, imperfecto o subóptimo.

Los daños son menoscabos sufridos tanto por los beneficiarios como por terceros ajenos al sistema. Siempre será necesario definir correctamente a cada grupo afectado, describiendo con claridad las características de cada cual. Esto ayudará a reconocer patrones presentes o cualidades que puedan estar sujetas a un mayor o menor escrutinio que el que se considera óptimo en el contexto donde se utiliza el sistema.

Ejemplo de grupos afectados

Hay que tener presente que los grupos afectados por los ADS no solo son aquellos que los utilizan o que se ven directamente implicados en las acciones o recomendaciones provenientes de tales sistemas. Muchas veces también existen terceros no considerados, ajenos a los mismos, que resultan afectados por su uso.

Tómese, por ejemplo, un sistema que determina la frecuencia del transporte público y cuyas decisiones afectan de forma clara y directa a sus usuarios en indicar si debería haber mayor frecuencia en un periodo del día para optimizar el uso de recursos públicos y la satisfacción de los usuarios. Sin embargo, hay otros grupos que también resultan afectados como pueden ser los usuarios de vehículos particulares, peatones y ciclistas, dado que sus tiempos de desplazamiento también serían afectados por esa frecuencia.

Como se indicó anteriormente en el recuadro sobre Criticidad de los Sistemas, los daños pueden incidir de forma grave en dos tipos de factores clave: (i) aquellos que directamente influyen en la asignación o restricción de derechos, ayudas o subsidios, y (ii) aquellos que son parte de una cadena de procesos y que, en caso de fallar o de que se produzca algún error, pueden afectar a cualquiera de los elementos que repercuten en la prestación de un servicio gubernamental. Es en esta segunda instancia cuando se presentan las principales amenazas de ataques cibernéticos que han afectado a algunos de los servicios del Estado.

La Unión Europea ha catalogado cuatro niveles de riesgo para los modelos de IA:

(i) **Riesgo inadmisible:** Aquellas aplicaciones que son nocivas para la salud e integridad de las personas y que contravienen derechos fundamentales. Están prohibidas.

(ii) **Alto riesgo:** Aplicaciones que tienen un impacto negativo en la seguridad de las personas o aquellas que son componentes de seguridad de sistemas mayores. Deberán ser evaluadas por terceros y cumplir con la regulación sectorial antes de entrar en funcionamiento.

(iii) **Riesgo limitado:** Aplicaciones que tienen un bajo nivel de riesgo pero que deben cumplir con los requerimientos de transparencia e información para con aquellos ciudadanos que están siendo sujetos de un tratamiento automatizado.

(iv) **Riesgo mínimo:** Cualquier sistema cuya aplicación no implique ningún riesgo. Los desarrolladores podrán acogerse a códigos de conducta en forma voluntaria.

Infortunadamente, cuando se trata de sistemas de ADS no siempre es posible obtener una explicación acerca de la razón y causalidad del error y su consiguiente daño. Esto es lo que se conoce como problema de “caja negra”. Por eso las auditorías técnicas se limitan a evaluar si se tomaron o no las precauciones necesarias al desarrollar el sistema.

Caja Negra

Se utiliza la metáfora de "caja negra" para hacer referencia a aquellos sistemas cuyo funcionamiento interno se desconoce, o bien porque es imposible entenderlo o porque hacerlo resulta demasiado costoso y por lo tanto poco razonable (por ejemplo, tratar de interpretar una red neuronal). En tales casos no es posible para un humano discernir la manera en que ciertos insumos (por ejemplo datos) llevan al sistema a arrojar un resultado (por ejemplo, una acción o recomendación determinada)a.

Si bien la utilización de estos modelos se puede justificar en razón de su mejor rendimiento, ciertamente se contrapone a la búsqueda de transparencia en la implementación de ADS.

[a] Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Disponible en: https://www.auditingalgorithms.net/

2.11 Perfiles y funciones de las persona que colaboran con la auditoría

En caso de requerir aclaraciones o más información sobre los antecedentes del sistema sujeto al proceso de auditoría, se deberá registrar a las personas de contacto relevantes con sus respectivas funciones en el organismo que lo implementa, ya que estas serán las responsables de suministrar la información solicitada. A continuación se describen los cargos y funciones que para estos sistemas podrían existir en una entidad promedio[20]:

» **Director de información** (CIO por sus siglas en inglés): Persona encargada de los sistemas informáticos y tecnológicos de una entidad. Decide y dirige los desarrollos tecnológicos para alcanzar los fines de la institución.

» **Director de privacidad** (CPO por sus siglas en inglés): Persona encargada de tomar las decisiones en materia de privacidad de la institución y de velar por la protección de los intereses de los beneficiarios en esta área.

» **Director de seguridad de la información** (CISO por sus siglas en inglés): Persona encargada de velar por la seguridad de la información que produce y posee la entidad.

» **Director jurídico:** Persona a cargo de los asuntos legales y de garantizar el cumplimiento normativo por parte de la institución.

» **Desarrollador de software:** Persona a cargo de programar el sistema y de convertir los requerimientos institucionales en un software que cumpla con los fines técnicos deseados.

» **Analista de datos:** Persona a cargo de analizar, ordenar y depurar los datos para que sirvan de insumo en la toma de decisiones dentro de la institución.

» **Ingeniero de datos:** Persona a cargo de construir y mantener las bases de datos y de prepararlas para que puedan ser posteriormente utilizadas por el analista de datos.

» **Director del proyecto:** Persona a cargo de que el proyecto se lleve a cabo, manteniendo su cohesión y distribuyendo las tareas en el equipo.

» **Dueño del producto** (Product Owner): Persona a cargo de las tareas prácticas del desarrollo del sistema relativas a estrategia, ejecución y lanzamiento.

» **Experto en el rubro:** Persona que posee el conocimiento acerca del campo de actividad que corresponde a la institución y que pueda poner en contexto las necesidades de los usuarios.

20 Habrá entidades de menor tamaño donde solo haya una persona a cargo, así como entidades de mayor envergadura que cuenten con equipos completos dedicados a cumplir las tares de una cierta función.

3. Etapas de una auditoría algorítmica

Si bien actualmente no existe un modelo único para conducir auditorías algorítmicas, en esta guía se utilizará el de Raji y Smart, et al. (2020), el cual comprende seis etapas: (i) definición del alcance de la auditoría, (ii) mapeo de las partes interesadas, (iii) recolección de documentación, (iv) realización de pruebas, (v) análisis de resultados, y (vi) post auditoría[21].

A continuación, se listan las tareas que habrán de realizarse en cada una de las seis etapas arriba mencionadas:

» Definición del alcance de la auditoría

- Recopilación del documento de requisitos del producto (PRD por sus siglas en inglés)
- Revisión de los principios considerados en el diseño del sistema
- Análisis de casos de uso similares
- Evaluación de impacto social por medio de la finalidad del ADS.

» Mapeo de partes interesadas

- Formulación de preguntas y entrevistas al equipo
- Transcripción y sistematización de respuestas

» Recolección de documentación

- Elaboración de listas de control con los puntos a auditar
- Elaboración de perfil de datos
- Elaboración de perfil del modelo

» Realización de pruebas

- Revisión de la documentación
- Simulación de fallas y búsqueda de vulnerabilidades
- Elaboración de la matriz de riesgo correspondiente al uso del sistema

21 Raji et al. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Disponible en https://arxiv.org/pdf/2001.00973.pdf

» Análisis de resultados

- Actualización y formalización de la matriz de riesgo
- Elaboración de plan de acción o de mitigación de riesgos
- Recopilación del detalle y evolución del desarrollo del sistema
- Informe de auditoría

Nótese que estos pasos no siempre serán secuenciales y que es posible que, desde las etapas tempranas de la auditoría se descubra que el sistema es inviable, lo cual hará innecesario avanzar hacia etapas posteriores.

En el anexo de este documento se incluyen preguntas guía para realizar correctamente una auditoría.

Hasta la fecha, las diversas regulaciones existentes en América Latina y el Caribe en materia de inteligencia artificial no hacen referencia directa a la responsabilidad algorítmica, a diferencia de países con jurisdicciones más maduras en el desarrollo de tales temas como son Canadá[22], Suecia[23] o el Reino Unido[24].

Debido al crecimiento en el número de iniciativas de ley que influyen en el desarrollo de los sistemas autónomos, así como de las políticas acerca del uso de la IA en los países de la región, la respuesta vendrá determinada por el cumplimiento tanto de la legislación y las políticas que buscan asegurar un uso adecuado del sistema, como de las normativas propias que regulen la industria en la que este se implemente. Es necesario tener en cuenta que existen áreas relacionadas que tendrán incidencia en la revisión de los estándares de cumplimiento, por ejemplo en materia de protección de datos, ciberseguridad, leyes antidiscriminación o incluso normativas sectoriales.

En lo concerniente al sector público, actualmente no hay una práctica estandarizada para realizar auditorías algorítmicas. Sin embargo, han surgido algunas iniciativas que apuntan a consolidar experiencias en diversos casos, áreas y jurisdicciones. A modo de ejemplo está el documento titulado "A White Paper for Public Auditors",[25] elaborado por las autoridades de auditoría de Finlandia, Alemania, los Países Bajos, Noruega y el Reino Unido, con base en su experiencia en la materia.

Con posterioridad a la auditoría, y a la luz de los resultados obtenidos, será necesario que la entidad determine si es posible continuar utilizando el sistema o si se debe modificar en parte o en su totalidad, de acuerdo con las respuestas obtenidas (clasificadas en el anexo, según su relevancia, en aquellas de suma urgencia, suma importancia o revisión recomendable). En caso de requerirse, se deberá implementar el plan de acción o de mitigación de riesgos, sumado a un posterior seguimiento continuo de su implementación.

22 A modo de ejemplo, en Canadá la Directiva sobre toma de decisiones automatizadas de 2019 tiene como finalidad reducir el riesgo que estos sistemas presentan y lograr decisiones administrativas más eficientes, precisas, consistentes e interpretables, en concordancia con la ley canadiense. Para ello se expande la implementación de auditorías, se facilita el acceso a información y se eleva el estándar en materia de calidad de datos. Directive on Automated Decision-Making, disponible en https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

23 Automated decision-making in public administration – effective and efficient, but inadequate control and follow-up, disponible en https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration---effective-and-efficient-but-inadequate-control-and-follow-up.html

24 Guidance on the AI auditing framework, Draft guidance for consultation, disponible en https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

25 Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Disponible en: https://www.auditingalgorithms.net/

4. Principios para los ADS

Con el fin de promover la implementación y el uso ético de sistemas basados en inteligencia artificial, diversas jurisdicciones y organismos han adoptado principios con los cuales se busca guiar su implementación, tanto en temas particulares como en su totalidad. Están, por ejemplo, los principios enunciados en el Artículo 5º del Reglamento General de Protección de Datos (RGPD) europeo[26], que regulan específicamente el procesamiento de datos personales. Para el sistema en su totalidad, están los principios consignados en el documento Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI[27] del Berkman Klein Center for Internet and Society de la Universidad de Harvard.

En esta guía se tendrá en cuenta el listado de principios éticos propuestos por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) en su documento Recommendation of the Council on Artificial Intelligence.[28] Se trata del primer conjunto de estándares de políticas intergubernamentales sobre inteligencia artificial integrado por los principios traducidos en el documento "Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe",[29] los cuales se resumen a continuación.

Crecimiento inclusivo, desarrollo sostenible y bienestar. Las partes interesadas deberán participar activamente en la gestión responsable de una IA que haya sido concebida para alcanzar resultados que beneficien a las personas y al planeta. Con el uso adecuado de la IA se podrá promover el aumento de las capacidades humanas y de la creatividad, la inclusión de poblaciones minoritarias, la reducción de las desigualdades económicas y sociales, así como la protección de entornos naturales, estimulando para ello el crecimiento inclusivo, el desarrollo sostenible y el bienestar.

Valores centrados en el ser humano y la equidad. Los actores del ecosistema de IA deben respetar el Estado de Derecho, los derechos humanos y

26 https://gdpr-info.eu/art-5-gdpr/

27 Fjeld y Nagy. Principled Artificial Intelligence. Disponible en: https://cyber.harvard.edu/publication/2020/principled-ai

28 OECD. Recommendation of the Council on Artificial Intelligence. Disponible en: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

29 BID, fAIr LAC. Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe. Disponible en: https://publications.iadb.org/publications/spanish/document/fAIr_LAC_Adopción_ética_y_responsable_de_la_inteligencia_artificial_en_América_Latina_y_el_Caribe_es.pdf

los valores democráticos a lo largo de todo su ciclo de vida. Entre estos últimos sobresalen la libertad, la dignidad y la autonomía; la privacidad y protección de los datos; la no discriminación y la igualdad; y la diversidad, la equidad, la justicia social y los derechos laborales internacionalmente reconocidos. Con este fin, los actores de la IA deben implementar mecanismos y salvaguardias de protección de derechos como el de la autodeterminación de los individuos. Estos deben ajustarse al contexto y ser consistentes con el estado del arte.

Transparencia y explicabilidad. Los ADS deben permitir que los actores del ecosistema entiendan su funcionamiento y posibles resultados. Por lo tanto, los sistemas implementados deberán regirse por los principios de transparencia y divulgación responsable y leal de la información.

Se deberá proporcionar información relevante tanto para quienes usan los sistemas como para quienes son sujetos pasivos del análisis. La información deberá ajustarse al contexto del receptor de la información, de manera tal que este sea capaz de entenderla completa y correctamente.

Los objetivos son: (i) fomentar una comprensión general acerca del funcionamiento de los sistemas de IA; (ii) procurar que las partes interesadas tomen plena conciencia de sus interacciones con tales sistemas; (iii) asegurarse de que los beneficiarios y sujetos pasivos entiendan los posibles resultados y riesgos de la utilización de los ADS, y (iv) permitir que las personas afectadas adversamente por un sistema de IA impugnen sus resultados basándose en información clara y fácil de entender sobre los factores y la lógica que sirvieron de base para la predicción, recomendación o decisión que se busca refutar.

Es fundamental que los responsables por la toma de decisiones también entiendan el funcionamiento y los posibles riesgos que entraña el uso de los ADS, a fin de incorporar análisis propios allí donde la máquina pueda fallar o presentar riesgos. Tal como se describe más adelante en la sección 5 de este documento sobre el caso de uso de COMPAS (Correctional Offender Management Profiling for Alternative Sanctions),[30] se buscaba identificar el riesgo de que las personas procesadas judicialmente volvieran a cometer delitos.

El uso de COMPAS generó un gran revuelo, ya que mostraba un sesgo favorable hacia personas de piel blanca y uno adverso hacia aquellas con piel más oscura. Algo parecido sucedía en el caso de delitos cometidos por hombres y mujeres, siendo estas últimas las más castigadas.

Dado que ninguno de los sujetos pasivos del sistema COMPAS conocía su funcionamiento por carecer de transparencia, los jueces que se apoyaban en él no cuestionaron sus recomendaciones. Esto solo ocurrió posteriormente, a partir de un reportaje publicado en prensa como resultado del cual se dejó de usar el sistema.

Si el sistema hubiera sido transparente, habría sido evidente que estaba tomando en consideración elementos que no son propios de una sanción judicial como por ejemplo el origen étnico, la composición familiar y/o la escolaridad de los imputados. De la misma manera, estos últimos habrían tenido la posibilidad de defenderse de las penas que con base en el ADS se les impusieron, pues evidentemente eran contrarias al debido proceso.

La transparencia de los sistemas no solo permite que los sujetos pasivos o receptores de sus acciones ejerzan sus derechos; también contribuye a que los responsables por la toma de decisiones ponderen y analicen la validez de la recomendación, de forma tal que la entiendan plenamente y determinen si constituye un elemento que respeta la dignidad de las personas, los derechos humanos y el Estado de Derecho.

Robustez, seguridad y protección. Estos son tres elementos esenciales en todo sistema de IA, por las siguientes razones:

» Los sistemas de IA deben ser robustos, seguros y protegidos durante todo su ciclo para que, en condiciones de uso normal, uso

30 Brennan, T. y Dieterich, W. Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). Disponible en https://www.researchgate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS.

previsible, uso incorrecto u otras condiciones adversas, funcionen adecuadamente y no supongan un riesgo poco razonable para la seguridad.

» Para ello, los actores de la IA deben garantizar la trazabilidad permanente de los conjuntos de datos, procesos y decisiones tomadas durante el ciclo de vida del sistema de IA. Así será posible analizar correctamente, y en consonancia con el estado del arte, sus resultados y respuestas a las preguntas que se le formulen.

» En función de sus labores, del contexto y de su capacidad de actuación, los actores de la IA deberán aplicar continuamente un enfoque sistemático de gestión de riesgos en cada fase del ciclo de vida del sistema. Esto con el fin de abordarlos de la mejor manera, incluyendo aquellos relativos a la privacidad, a la seguridad digital y a los posibles sesgos.

Rendición de cuentas. Los actores de la IA deben ser responsables del buen funcionamiento de tales sistemas y del respeto por los principios antes mencionados, en función de sus deberes, del contexto y del estado del arte de la tecnología.

5. Caso de uso: ADS en la vigilancia predictiva

Para poner de manifiesto la relevancia de la realización de una auditoría algorítmica, a continuación se analiza su utilización en el contexto de la **vigilancia policial predictiva**.

En Estados Unidos se ha implementado el uso de ADS en el análisis de riesgo para la comunidad constituida por los imputados de delitos en distintos estados. El sistema utilizado, desarrollado por la empresa Northpointe, recibe el nombre de COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). Este sistema suministra un puntaje a la corte respectiva con base en las respuestas a un cuestionario compuesto por 137 preguntas del siguiente tenor: ¿Tu padre/madre estuvo en prisión alguna vez? ¿Qué tan seguido te metías en peleas en el colegio? Estas preguntas eran o bien respondidas por los imputados o se obtenían de sus historiales delictivos.

El año 2013, Eric Loomis fue arrestado por conducir un vehículo en el que iban personas que habían participado recientemente en un tiroteo. Como resultado de la recomendación hecha por el sistema COMPAS, que lo calificó como un sujeto altamente peligroso para la comunidad, fue sentenciado a seis años en prisión y a cinco años de supervisión ampliada[31].

En este caso saltan a la luz alertas en distintas materias: respecto a la idoneidad del uso del sistema en el contexto mencionado, a su precisión y a los sesgos que puedan contener los datos utilizados, entre otras. De ahí que, antes de proceder a utilizar un ADS en contextos tan complejos como la administración de justicia, es necesario responder siempre a las preguntas que a continuación se formulan, y que se ejemplificarán con el caso de COMPAS

¿Se ha definido un propósito claro para el uso del sistema?

¿De qué manera se asegura que el sistema no se utilice con fines distintos a aquellos para los cuales fue desarrollado?

Respecto a estos interrogantes, en el caso de COMPAS Tim Brennan, fundador de Northpoint,

31 Wisconsin Supreme Court. *State v. Loomis.* Disponible en: https://harvardlawreview.org/2017/03/state-v-loomis/

señaló que su enfoque al diseñar este sistema era reducir el crimen, no determinar penas. Como se describió anteriormente para este caso de uso, el sistema se degeneró y terminó utilizándose como fundamento para la determinación de la culpabilidad del imputado, lo cual dista del objetivo que se tuvo originalmente para su desarrollo.

¿Se ha probado el sistema en diferentes grupos demográficos para mitigar los sesgos existentes?

¿Se han tomado medidas para mitigar los sesgos históricos en las bases de datos utilizadas?

El hecho de que el cuestionario de COMPAS estuviera compuesto por preguntas acerca de la infancia, antepasados o barrio en el que residía el imputado, debió prender las alarmas acerca de los sesgos que se pudieran introducir en los datos. En este caso no se tomaron las medidas apropiadas para mitigar los sesgos históricos existentes en los datos utilizados, lo cual tuvo como consecuencia una asignación errónea de puntaje de riesgo a individuos con registros de antecedentes dispares.

¿La definición de la arquitectura y de las técnicas utilizadas concuerdan con las necesidades de transparencia y explicabilidad de las decisiones que exige el sector de actividad en el que se inserta el sistema?

Existen ciertos sectores en los que la explicabilidad de las decisiones es fundamental para una adecuada aceptación de los sistemas de IA por parte de la sociedad. En materia de administración de justicia, la explicabilidad es un imperativo. En el caso arriba mencionado, las partes interesadas no tenían conocimiento acerca de cómo se asignaba el puntaje, ya que Northpointe sostenía que esa información era un secreto comercial sujeto a reserva.

Considerando la inestabilidad que caracteriza a varios modelos de aprendizaje automático, ¿se ha validado el modelo en múltiples ocasiones y escenarios con el fin de cerciorarse de que el sistema responda correctamente en distintos contextos?

¿Cómo se han definido los puntos óptimos de sensibilidad y especificidad en la curva de ROC?[32] ¿Son adecuados para la industria en la que se busca implementar el sistema?

Una evaluación del sistema realizada con posterioridad, en la cual se analizaron 16 mil casos, reveló que su precisión era cercana al 71%.[33] Al tratarse de un contexto de implementación bastante sensible como lo es la administración de justicia, es evidente que la precisión revelada dista bastante de lo que podría considerarse idóneo. De lo anterior se concluye que hicieron falta mayores periodos de prueba y que se requerían parámetros de validación más elevados.

32 La curva de ROC (característica operativa del receptor por sus siglas en inglés) es una herramienta estadística que permite evaluar la precisión de las predicciones de un modelo. Por ejemplo, si se intenta implementar un modelo que clasifica a personas según su riesgo de cometer delitos, la curva de ROC puede evaluar a precisión de dicho modelo.

33 Angwin et al. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

6. Comentarios finales

La inteligencia artificial desempeña un papel fundamental en nuestro día a día y en nuestra convivencia como sociedad, hasta tal punto que se hace cada vez más difícil pensar en siquiera una instancia en que no interactuemos con sistemas inteligentes en el mundo de hoy. Dispositivos móviles, electrodomésticos, medios de transporte, entre tantos otros, descansan en este tipo de sistemas para hacer que nuestras actividades sean más sencillas, cómodas y seguras.

Al igual que la tecnología de IA, el campo de la auditoría algorítmica avanza a un ritmo acelerado, con lo cual crece la importancia de su utilización. Se trata de un trabajo en continua evolución, cuyo alcance está en flujo permanente. Así pues, sus contenidos deberán ser actualizados de manera regular conforme al desarrollo de nuevas herramientas tecnológicas y a la regulación correspondiente.

Dada la masificación de los ADS en la sociedad, y particularmente en áreas donde su uso puede exigir precauciones extraordinarias, es necesario realizar revisiones constantes que guíen su implementación correcta. Ya se han registrado casos de sistemas que, por fallas de diseño o desarrollo, generan impactos adversos importantes en nuestras vidas y en la sociedad: estos van desde el aumento de tarifas en el transporte público, hasta sentencias injustas.

Dentro de los desafíos que esta área conlleva para el sector público figura el tránsito de la realización de auditorías algorítmicas como mecanismo voluntario, hacia su inclusión como parte de una política estructurada sobre la materia, o bien como parte de una regulación de amplio alcance en materia de responsabilidad algorítmica.

Esta guía no pretende ser solo un instrumento práctico que contribuya a la vigilancia de áreas críticas y a la mitigación de riesgos o peligros que quizás no sean patentes a simple vista. También se espera que sirva como instrumento que ayude en tomar conciencia acerca de las implicaciones y consecuencias de la puesta en marcha de un ADS. Esperamos que todos los equipos, tanto los de las entidades públicas como los de desarrolladores de ADS, sean consciente de la relevancia de su trabajo. Si queremos garantizar un futuro más justo y seguro, es necesario entender y a hacer que se entienda plenamente su pertinencia.

Referencias

Ada Lovelace Institute (2021). Algorithmic Accountability for the Public Sector. Disponible en: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

---------- (2020). Examining the Black Box: Tools for assessing algorithmic systems. Disponible en https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Angwin, J., Larsson, J., Mattu, S. y Krichner L. (2016). Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

BID (2021) Uso responsable de IA para política pública: manual de formulación de proyectos. Disponible en https://publications.iadb.org/es/uso-responsable-de-ia-para-politica-publica-manual-de-formulacion-de-proyectos

BID Lab (2021). Guía de aplicación. Autoevaluación ética de IA para actores del ecosistema emprendedor. Disponible en: https://publications.iadb.org/publications/spanish/document/Autoevaluacion-etica-de-IA-para-actores-del-ecosistema-emprendedor-Guia-de-aplicacion.pdf

BID/ fAIr LAC (2020). Adopción ética y responsable de la Inteligencia Artificial en América Latina y el Caribe. Disponible en: https://publications.iadb.org/publications/spanish/document/fAIr_LAC_Adopción_ética_y_responsable_de_la_inteligencia_artificial_en_América_Latina_y_el_Caribe_es.pdf

Brennan, T. y Dieterich, W. (2017). Correctional Offender Management Profiles for Alternative Sanctions (COMPAS). Disponible en https://www.researchgate.net/publication/321528262_Correctional_Offender_Management_Profiles_for_Alternative_Sanctions_COMPAS.

Fjeld, J. y Nagy, A. (2020). Principled Artificial Intelligence. Disponible en: https://cyber.harvard.edu/publication/2020/principled-ai

Government of Canada. (2021). Directive on Automated Decision-Making, disponible en https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

IA Now Institute (2028). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. Disponible en: https://ainowinstitute.org/aiareport2018.pdf

Intersoft Consulting. S.f. Data Protection Regulation. Disponible en: https://gdpr-info.eu/art-5-gdpr/

INTOSAI (2019). Performance Audit Principles. Disponible en https://www.intosai.org/fileadmin/downloads/documents/open_access/ISSAI_100_to_400/issai_300/ISSAI_300_en_2019.pdf

ISO 19011-2018 (2018). Disponible en https://www.iso.org/standard/70017.html

Moss, E., Watkins, E. A., Singh, R., Elish, M. C. y Metcalf, J. (2021). Assembling Accountability: Algorithmic Impact Assessment for the Public Interest. Disponible en https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

OECD (2019). Recommendation of the Council on Artificial Intelligence. Disponible en https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

OECD/BID (2020). IA Responsable: Manual técnico: Ciclo de vida de la inteligencia artificial. Disponible en https://publications.iadb.org/publications/spanish/document/IA-Responsable-Manual-tecnico-Ciclo-de-vida-de-la-inteligencia-artificial.pdf

Raji, I. D., et al. (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. Disponible en https://arxiv.org/pdf/2001.00973.pdf

Supreme Audit Institutions of Finland, Germany the Netherlands, Norway, and the UK. Auditing Machine Learning Algorithms, A white paper for public auditors. Disponible en: https://www.auditingalgorithms.net

Swedish National Audit Office (2020). Automated decision-making in public administration – effective and efficient, but inadequate control and follow-up. Disponible en https://www.riksrevisionen.se/en/audit-reports/audit-reports/2020/automated-decision-making-in-public-administration---effective-and-efficient-but-inadequate-control-and-follow-up.html

UK Information Commissioner's Office (2020). Guidance on the AI auditing framework, Draft guidance for consultation. Disponible en https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

US Congress (2019). Algorithmic Accountability Act of 2019.


BID
Mejorando vidas