Digital Future Society

Hacia un uso responsable de los algoritmos: métodos y herramientas para su auditoría y evaluación

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

Mobile WorldCapital Barcelona

Sobre Digital Future Society

Digital Future Society es una iniciativa transnacional sin ánimo de lucro que conecta a responsables políticos, organizaciones cívicas, expertos académicos y empresarios para explorar, experimentar y explicar cómo se pueden diseñar, usar y gobernar las tecnologías a fin de crear las condiciones adecuadas para una sociedad más inclusiva y equitativa.

Nuestro objetivo es ayudar a los responsables políticos a identificar, comprender y priorizar los desafíos y las oportunidades fundamentales, ahora y en los próximos diez años, en relación con temas clave que incluyen la innovación pública, la confianza digital y el crecimiento equitativo.

Para más información, visite digitalfuturesociety.com

Un programa de

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

Mobile WorldCapital Barcelona

Publicado
Febrero del 2024

Digital Future Society

Contenidos

Seis ideas clave

1. **La evaluación de algoritmos es una cuestión necesaria y crítica.**

   Los sistemas algorítmicos están produciendo impactos en la sociedad, por ejemplo, generando casos de discriminación y sesgos, efectos negativos para la sostenibilidad ambiental o vulneraciones de la privacidad, entre otros. Es indispensable, pues, analizar el uso de algoritmos desde una perspectiva holística, para detectar problemas y ofrecer medidas de mitigación de riesgos que fortalezcan una innovación sostenible.

---

2. **La evaluación de algoritmos requiere atender diferentes dimensiones.**

   En función de una serie de dimensiones (foco, locus, actores, momento, tema, ámbito, etc.), las evaluaciones de algoritmos pueden centrarse de forma prioritaria en diferentes aspectos del proceso, incluyendo los tecnológicos, pero también otros más humanos, que consideren la interrelación de la tecnología con el contexto social en el que se despliega.

---

3. **No existe una receta única para evaluar algoritmos.**

   Los métodos disponibles varían de acuerdo con los enfoques, y tienen ventajas y limitaciones en determinados contextos. La intersección entre métodos (auditorías de código, *scraping*, *checklists*, estudios de caso, etc.) y las dimensiones de las evaluaciones algorítmicas ofrece un mapa de situación sobre sus oportunidades y limitaciones.

---

4. **Las evaluaciones de algoritmos no se desarrollan en espacios aislados.**

   En el momento de analizar el funcionamiento y los efectos de un algoritmo, se debe tener en cuenta el ecosistema de actores que entran en juego en estos sistemas. Se pueden identificar tres niveles de gobernanza: interacción entre los ámbitos público y privado y el tercer sector (*macro*); sectores de actividad, como la salud, la educación, la seguridad, etc. (*mezzo*); y actores que diseñan, implementan, usan y auditan los algoritmos (*micro*). Comprender estas dinámicas permitirá evaluar algoritmos de una forma adecuada y lograr una rendición de cuentas más completa.

5. Las personas siempre deben estar en el centro.

Independientemente del enfoque que se siga y los métodos que se usen, es importante dar protagonismo a las personas. Eso significa que se debe hacer un esfuerzo para entender el funcionamiento de los algoritmos y el impacto que tienen en la vida, sobre todo, de las poblaciones en situación de exclusión y vulnerabilidad. También se ha de atender a las personas que diseñan e implementan estos sistemas, a su relación con las estructuras organizativas y al contexto social más amplio en el que se desenvuelven.

---

6. Debe darse importancia a los estándares y a los organismos supervisores de algoritmos.

Los ámbitos público y privado y el tercer sector deben trabajar conjuntamente para definir unos estándares claros en los procesos de evaluación algorítmica. Así se podrán evitar prácticas no adecuadas que resten validez a estos procesos, al tiempo que se impulsan criterios compartidos para avanzar hacia mejores evaluaciones de algoritmos. La existencia de organismos de supervisión de algoritmos también contribuirá decididamente a la eficacia y confianza en estos procesos, sobre todo si se promueve la cooperación internacional y el intercambio de aprendizajes.

Introducción

A medida que avanza la implementación de algoritmos en una variedad de contextos y sectores, se vuelve cada vez más necesario el debate sobre los aspectos éticos y de gobernanza de la inteligencia artificial (IA). Más allá de las promesas sobre el aumento de la eficiencia y la eficacia, los sistemas algorítmicos pueden contener sesgos o tomar decisiones erróneas que generan efectos indeseados en la vida de las personas. Ante este panorama, se ha considerado que las evaluaciones de algoritmos pueden contribuir a paliar dichos efectos, con la detección de aspectos problemáticos como la discriminación de determinadas poblaciones, la distorsión de la realidad o la explotación de información personal (Bandy 2021). Concretamente, los procesos de evaluación impulsarían el cumplimiento de los principios éticos que se recogen en normativas y documentos estratégicos de la IA.

En los últimos años, se han publicado una gran cantidad de artículos académicos e informes que buscan dilucidar los detalles que deben incluirse en una evaluación algorítmica. Es decir, desean responder en términos prácticos a la siguiente pregunta: **¿cómo se pueden evaluar los algoritmos para detectar los potenciales problemas que contienen y/o que se derivan de su uso, así como contribuir a su mitigación?** Los enfoques son variados. En algunos casos, se da prioridad al análisis del sistema algorítmico desde un punto de vista eminentemente tecnológico; en otros, se aboga por el estudio de riesgos e impactos en la población y en las organizaciones, con una perspectiva más general y holística. También es posible desarrollar estos procesos antes o después de la implementación del sistema, y con la participación de actores externos o bien sin ella. Las opciones, en este sentido, son variadas.

Con el fin de aportar claridad sobre esta temática compleja, en este documento se exploran y sistematizan las alternativas que existen en el desarrollo de evaluaciones de algoritmos, y se ofrece una panorámica de los métodos y herramientas que se pueden usar para evaluar algoritmos en función de los distintos objetivos planteados y recursos disponibles. También se explica el ecosistema de actores y sectores involucrados en este tipo de procesos, considerando el marco más general de la rendición de cuentas algorítmica. Con base en todos estos contenidos, se presentan, finalmente, seis recomendaciones para mejorar las evaluaciones de algoritmos en el futuro.

Para el desarrollo de este informe, se siguió una **metodología** cualitativa, compuesta por tres fases (véase el Anexo):

- **Revisión sistemática de literatura académica.**
  En este caso, se usó ASReview, una herramienta de aprendizaje activo (*active learning*) que facilita la selección de artículos relevantes a través del entrenamiento de modelos de aprendizaje automático (*machine learning*). En esta fase, de una muestra de casi 3.000 artículos, se analizó el contenido de un total de 64 documentos. Con ello se han sintetizado los resultados de investigación más recientes sobre este tema a nivel internacional.

- **Análisis documental de literatura gris.**
  Después, se realizó una búsqueda de informes y otros tipos de publicaciones, distintas a los artículos académicos, elaboradas y difundidas por organismos públicos, organizaciones del tercer sector, universidades, *think tanks* y otras entidades. En este caso, se identificaron 60 documentos que fueron incluidos en la síntesis de resultados. El objetivo fue conocer también trabajos de carácter más aplicado sobre el tema, ampliando el foco a un abanico lo más amplio posible de fuentes documentales.

- **Realización de entrevistas semiestructuradas a especialistas en IA y evaluaciones de algoritmos.**
  En esta fase, se llevaron a cabo entrevistas a quince especialistas que trabajan en organismos internacionales, organismos europeos, empresas y consultoras privadas, universidades y organizaciones del tercer sector, nueve de ellas en inglés y seis en español. El propósito aquí fue completar el trabajo de campo con una aproximación a las percepciones de personas expertas, con la finalidad de identificar aspectos o matizar cuestiones menos atendidas en los artículos académicos o documentos aplicados.

A partir de estas fuentes de información, se ha avanzado hacia la profundización en un tema de gran relevancia en las sociedades actuales, que debe ser de especial interés para representantes políticos, personal de organismos públicos, organizaciones del tercer sector, activistas y especialistas en la materia, así como para el público general. El debate sobre los procesos de evaluación de algoritmos invita a reflexionar sobre el impacto de la IA en las personas y organizaciones, y sobre el futuro que se desea construir en torno a la relación entre personas y máquinas.

1. Razones para evaluar algoritmos en el contexto actual de la inteligencia artificial

El uso de algoritmos y sistemas de IA está generando reacciones dispares y, a menudo, contrapuestas dentro de sectores de actividad cada vez más diversos. Tanto en los discursos presentes en la agenda pública y política como en los medios de comunicación o en estudios realizados sobre las percepciones individuales, se pueden encontrar posturas pesimistas y optimistas sobre la IA y el futuro de la humanidad. Se señala, por ejemplo, que estos sistemas contribuirán con eficiencia al progreso humano o, por el contrario, que serán responsables de transformaciones negativas para la sociedad (Stanford University Human-Centered Artificial Intelligence 2023).

En línea con otras olas de innovación tecnológica previas y desde una perspectiva no determinista, el punto de partida de este documento es que no es posible predecir de forma exacta lo que ocurrirá en el futuro. Ahora bien, corresponde analizar cómo están funcionando los sistemas de IA basados en algoritmos, qué impacto real están produciendo y qué riesgos conllevan. Particularmente, en esta sección se intenta responder a la pregunta sobre la necesidad de evaluar algoritmos en un momento en que existe un cierto consenso sobre el inicio de la llamada Cuarta Revolución Industrial.

El campo de las evaluaciones de algoritmos ofrece algunas respuestas conceptuales y prácticas a estas inquietudes. En los últimos años, ha ido creciendo el debate en círculos académicos, organizaciones de la sociedad civil y gobiernos sobre la relevancia de estas evaluaciones en el contexto actual, y sobre las metodologías y herramientas más adecuadas para ejecutarlas. Pero, ¿por qué es importante desarrollar evaluaciones algorítmicas en organizaciones del sector público y privado? A continuación, se presentan algunos argumentos.

- **El impacto de la IA y del uso de algoritmos es algo tangible.**
  Es cierto que aún no se puede saber a ciencia cierta cómo será el futuro con la IA, pero sí existen evidencias de sus impactos negativos potenciales y actuales sobre determinadas poblaciones. Por ejemplo, los datos y modelos pueden contener sesgos de género, raza y otros que deriven en discriminación de algunos colectivos (Buolamwini y Gebru 2018; Morondo y Eguiluz 2022). También hay casos de algoritmos que se han usado para automatizar procesos de forma inadecuada, con consecuencias negativas para personas en situación de vulnerabilidad: entre ellas, la exclusión de determinados servicios públicos y de la asistencia social (Eubanks 2019).

  Por otro lado, se ha señalado que los sistemas de IA tienen un alto impacto ambiental, tanto por el hardware y los servidores necesarios como por el consumo ingente de recursos (materias primas, electricidad, etc.) y el tiempo de almacenamiento en la nube (Strubell et al. 2019). Es decir, hay razones de peso para evaluar los posibles efectos negativos que ya se están evidenciando en el diseño e implementación de algoritmos.

- **La evaluación es algo en parte nuevo, pero no tan nuevo.**
  Las evaluaciones de algoritmos se asientan en una amplia tradición de procesos de auditoría y evaluación en otros ámbitos. Está el caso, por ejemplo, de las auditorías financieras, pero también existen modelos para realizar evaluaciones del impacto social y ético, y del impacto en la privacidad, en la protección de datos y en los derechos humanos (Mantelero 2018). Estos marcos de análisis ofrecen una hoja de ruta valiosa para examinar el funcionamiento y los efectos de los sistemas algorítmicos en la sociedad, pero también se deben efectuar ajustes.

  Ante la complejidad de la IA y su creciente implementación en una gran variedad de escenarios, se hace indispensable contar con estrategias específicas para evaluar las particularidades de estos sistemas. Hay que prestar atención, además, a las limitaciones que se han identificado en sus procesos, como la dificultad para identificar determinados daños y el riesgo de que la evaluación se reduzca a comprobar una lista de indicadores sin ofrecer una reflexión más profunda (Mökander et al. 2022). De todos modos, con la perspectiva adecuada y aprovechando las lecciones aprendidas de otros sectores, las evaluaciones de algoritmos tienen un gran potencial para identificar y mitigar los impactos negativos de dichos sistemas.

- **Evaluar el uso de algoritmos es una creciente obligación legal y ética.**
  Salvo algunas excepciones —como la Ley Local 144 de la ciudad de Nueva York o la Directiva sobre Decisiones Automatizadas del Gobierno de Canadá—, por el momento existe escasa regulación relacionada con las evaluaciones de algoritmos. Se pueden encontrar, además, diferencias importantes en los modelos de la Unión Europea, Norteamérica y China, que probablemente lleven a desarrollos diferenciados en esta materia. Así, en la UE se apuesta por un enfoque más regulatorio, que se evidencia

en la Ley de IA que próximamente entrará en vigor[1]. Por su parte, en el ámbito anglosajón hay una tendencia a evitar las regulaciones excesivas, por lo que las evaluaciones de algoritmos probablemente tomen la forma de certificaciones *ex post* o códigos de conducta[2]. En el caso de China, el rol determinante del Estado y una cultura social anclada en el confucionismo podrían impulsar la ejecución de este tipo de procesos bajo la mirada atenta del Gobierno[3], con un papel secundario de otros actores.

Junto con esta aproximación, y teniendo en cuenta un enfoque centrado en países democráticos, para Ricardo Baeza-Yates, director de investigación del Institute for Experiential AI de la Northeastern University, de Estados Unidos, existe una distinción clara entre los países con mayor confianza en las instituciones y aquellos en los que no se percibe que tengan un buen funcionamiento. En el primer caso, podría apuntarse a un modelo que privilegia la rendición de cuentas (es decir, realizar evaluaciones para establecer responsabilidades una vez que se implementan los algoritmos), mientras que, en el segundo, tendría más peso la transparencia (es decir, hacer pública determinada información a lo largo de todo el ciclo de vida de la IA)[4].

Pese a esa diferencia de modelos, se considera que las evaluaciones son mecanismos indispensables para la rendición de cuentas algorítmica (Basu et al. 2021). Por lo tanto, independientemente de la forma que adopten (sea a través de regulaciones estrictas o de mecanismos voluntarios), se espera que los gobiernos y/o la sociedad civil empiecen a demandar cada vez más la ejecución de evaluaciones algorítmicas.

- **La evaluación algorítmica es una cuestión sociotécnica.**
  La creciente presencia de sistemas algorítmicos en diferentes esferas de la vida cotidiana requiere centrar la atención en sus implicaciones traspasando las fronteras de lo meramente tecnológico, para llegar a las preocupaciones de la agenda pública y social en términos más amplios. Se trata de un asunto relevante si se consideran los diferentes enfoques que existen en torno a la dimensión ética del uso de algoritmos, especialmente en relación con la equidad y la no discriminación. Por ejemplo, en muchos casos, desde el sector tecnológico, se ha entendido que los esfuerzos deben centrarse en la identificación y mitigación de sesgos en los datos y modelos, con un enfoque eminentemente técnico que no necesariamente aborda la raíz de la discriminación estructural (Morondo y Eguiluz 2022, p. 27).

[1] A finales del 2023 se alcanzó un acuerdo provisional entre la Presidencia del Consejo y el Parlamento Europeo: (https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence).

[2] Se debe considerar, de todos modos, que en el 2022 se introdujo en el Senado estadounidense la propuesta de Ley de Rendición de Cuentas Algorítmica (Algorithmic Accountability Act), que requeriría a las grandes compañías el desarrollo de evaluaciones de impacto de sus algoritmos (https://www.congress.gov/bill/117th-congress/senate-bill/3572).

[3] El Gobierno de China puso en marcha un registro obligatorio de algoritmos de recomendación, en el que las empresas, además de incluir información general sobre el sistema, deben subir un documento con una autoevaluación de seguridad. Los criterios para entender los riesgos de seguridad recaen únicamente en el Gobierno y los procesos de evaluación no están abiertos a la ciudadanía (Sheehan y Du 2022).

[4] Tanto la transparencia como la rendición de cuentas están interrelacionadas, pero en algunos casos se da más relevancia a una de ellas, si bien es cierto que las exigencias de transparencia son un elemento básico que está muy presente en todos los países.

Es decir, lo que se entiende por sesgos y discriminación en el mundo tecnológico no necesariamente coincide con las conceptualizaciones que se usan en otras disciplinas, a las que se debe prestar especial atención. Si se consideran los avances de la IA y su impacto en la vida cotidiana de cada vez un mayor número de personas, se debe trascender esta mirada estrictamente técnica para considerar la interdependencia de los aspectos tecnológicos con las dimensiones más humanas y sociales[5]. Esta perspectiva sociotécnica invita a añadir una mayor complejidad de las evaluaciones de algoritmos y un enfoque holístico que sea capaz de atender una realidad poliédrica en la que las personas son el eje central.

[5] Para Javier de la Cueva, patrono de la Fundación Civio y especialista en Derecho y Tecnologías de la Información y la Comunicación, es imposible separar la dimensión técnica de la política, social o cultural. En otras palabras, determinados constructos sociales, políticos o culturales se manifiestan en la forma en la que se diseñan e implementan las tecnologías.

2. Marco conceptual: definiciones y dimensiones de las evaluaciones de algoritmos

Una vez que se ha explicado la importancia de evaluar los algoritmos, es necesario realizar algunas precisiones conceptuales. Por ejemplo, ¿qué es exactamente una evaluación algorítmica? Encontrar una definición única puede ser una tarea complicada. Tanto en estudios académicos como en informes de organizaciones de la sociedad civil y organismos gubernamentales, se puede encontrar una variedad de enfoques de estos procesos.

Por lo general, se definen las auditorías o evaluaciones de impacto como mecanismos que permiten identificar comportamientos problemáticos de los sistemas algorítmicos (Bandy 2021), pero el acento se pone en distintos objetivos. Por ejemplo, en la detección de sesgos y discriminación en las decisiones algorítmicas[6] (Minkkinen et al. 2022; Sandvig et al. 2014); la evaluación de daños potenciales (Baykurt 2022); el análisis de los niveles de riesgo en términos de derechos humanos, ética y privacidad (Yam y Skorburg 2021), o el estudio del impacto en los derechos e intereses de determinados colectivos (Brown et al. 2021). En algunos casos, se considera que no solo se deben identificar los problemas, sino que también se deben apuntar las posibles soluciones y estrategias de mitigación[7].

[6] En este caso, es importante considerar los diferentes enfoques que existen para abordar la discriminación y los sesgos algorítmicos. Como destacan Morondo y Eguiluz (2022), un sistema algorítmico puede contener sesgos en los datos y modelos, que pueden alterar su funcionamiento y que pueden impulsar la aplicación de medidas de mitigación eminentemente tecnológicas. Esta perspectiva, sin embargo, no necesariamente tiene en cuenta la discriminación estructural que sufren algunas poblaciones. En este sentido, es importante avanzar hacia un enfoque holístico y multidisciplinar, que entienda los sesgos y la discriminación algorítmica en toda su complejidad.

[7] Adriano Soares Koshiyama, cofundador de la empresa Holistic AI, considera que las personas especialistas en evaluaciones algorítmicas deben desempeñar su función a la manera del personal sanitario: más allá del diagnóstico, se deben ofrecer también soluciones.

Una cuestión de términos

Entendiendo la diversidad conceptual que existe en torno a las evaluaciones de algoritmos (Ada Lovelace Institute 2020), en este documento se utiliza el término evaluación para hacer referencia de forma general a los procesos de análisis de los sistemas algorítmicos y la identificación de sus aspectos problemáticos. En este caso, realizamos la distinción entre esta idea paraguas de *evaluación* y el concepto de *evaluación de impacto*, que se considera una metodología específica para evaluar algoritmos y que se explicará más adelante (Ibid.).

Es importante destacar que, en este ámbito, el inglés ofrece una mayor riqueza terminológica para hacer distinciones conceptuales. Podríamos, en este sentido, establecer una diferencia clara entre *evaluation* (evaluación en sentido amplio), *impact assessment* (evaluación de impacto como metodología concreta) y *audit* (auditoría, centrada en el cumplimiento de requisitos técnicos o no técnicos). Junto a los anteriores, también se podrían mencionar otros conceptos relacionados, tales como *algorithmic accountability* (rendición de cuentas algorítmica), que aparecerán más adelante.

Más allá de esta conceptualización amplia, es importante realizar algunas distinciones. No todas las evaluaciones de algoritmos se desarrollan de la misma forma ni consideran los mismos elementos. A partir del estudio de la literatura académica y de la revisión documental de fuentes oficiales gubernamentales y de organismos dedicados a las auditorías y evaluaciones, así como de las entrevistas realizadas a personas expertas, se propone a continuación una tipología de evaluaciones de algoritmos en torno a diez dimensiones: foco, locus, actores promotores, rol de actores externos, momento, orientación hacia la normativa, temática, ámbito, nivel de acceso y metodología. La Tabla 1 resume dicha tipología e incluye, además, una serie de preguntas que permiten entender de forma más clara de qué trata cada una de las dimensiones.

Tabla 1.
Tipología de dimensiones en las evaluaciones de algoritmos

Dimensión	Categorías	Preguntas de referencia
Foco	Enfoque técnico Enfoque holístico	¿Existe acceso, especialmente, a los datos de entrenamiento, el modelo, las salidas de datos (*outputs*) y otros aspectos técnicos del sistema? ¿Se tiene acceso a información que permita analizar la relación entre el aspecto tecnológico y los aspectos sociales, organizativos, culturales, contextuales, etc.?
Locus	Interna Externa	¿La organización que implementa el sistema algorítmico está involucrada en el proceso de evaluación? ¿Se dispone de autorización expresa y acceso a la información interna sobre el desarrollo del algoritmo?
Actores promotores	Primarios (*first-party*) Secundarios (*second-party*) Terceros (*third-party*)	¿Quiénes lideran el proceso de evaluación algorítmica?
Rol de actores externos	Participativa No participativa	¿Qué actores participan en el proceso de evaluación del algoritmo? ¿Existe implicación de personas usuarias y de colectivos potencialmente afectados por el sistema algorítmico? En caso de que se dé la implicación de usuarios/as y personas afectadas, ¿de qué forma se desarrolla esa participación?
Momento	*Ex ante* *Ex post*	¿El proceso de evaluación del algoritmo se desarrolla antes o después de la implementación del sistema?

Dimensión	Categorías	Preguntas de referencia
Orientación hacia la normativa	Obligaciones de ley Cumplimiento de marco normativo Buenas prácticas Certificaciones	¿Cuál es la motivación para desarrollar la evaluación del algoritmo? ¿Se busca cumplir una normativa específica, impulsar buenas prácticas de forma voluntaria, obtener una certificación de calidad, etc.?
Tema	Uso de datos Ética y derechos humanos Gobernanza	¿Cuál es el tema central del proceso de evaluación? ¿Se da prioridad al análisis del uso de los datos, a aspectos de ética y derechos humanos, o a la gobernanza del sistema?
Ámbito	Aspecto concreto de un algoritmo Sistema completo	¿Cuál es el objeto de análisis? ¿Se trata de un elemento específico (datos, modelo, etc.) o se busca entender la totalidad del sistema?
Nivel de acceso	*White-box* *Black-box* Intermedio	¿Qué nivel de acceso tiene el equipo evaluador al sistema algorítmico? ¿Existe un acceso sin restricciones a toda la información o hay limitaciones para obtener datos internos?
Metodología	Auditorías Evaluaciones de impacto	¿Qué metodología específica se sigue en el proceso de evaluación? ¿Se busca analizar el algoritmo en función de una serie de criterios específicos, o se pretende entender sus potenciales riesgos o impactos?

Fuente de los datos: elaboración propia a partir de Ada Lovelace Institute 2020; Costanza-Chock et al. 2022; Meßmer y Degeling 2023; Metcalf et al. 2021; Kelly-Lyth y Thomas 2023; Koshiyama et al. 2021 y entrevistas a personas expertas

Foco

Una primera dimensión clave para entender los procesos de evaluación de algoritmos es el foco, que se refiere a su posible enfoque técnico u holístico. Cuando se sigue un enfoque eminentemente **técnico**, se busca entender el funcionamiento del algoritmo y/o los códigos que permiten transformar las entradas de datos (*inputs*) en salidas (*outputs*). Se evalúan los resultados y las decisiones tomadas por el algoritmo en función de unos criterios específicos y técnicos. Por ejemplo, se puede determinar si existen sesgos en los datos y modelos de los algoritmos.

Un caso paradigmático es la investigación de Buolamwini y Gebru (2018), en la que se auditaron sistemas de clasificación y se encontraron sesgos de género y raza. Los resultados se publicaron en el conocido artículo *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* (2018).

Por su parte, el enfoque **holístico** se centra en la detección de problemas desde una perspectiva más amplia, lo que implica trascender el ámbito específico del funcionamiento algorítmico para comprender el contexto, las estructuras, los actores y otros elementos que se interconectan con el despliegue de los algoritmos y que influyen en sus resultados. Pese a que se plantea una distinción clara, también es posible asumir que lo ideal es incorporar ambos enfoques para que el proceso de evaluación sea lo más completo posible.

Un ejemplo interesante que está a medio camino entre ambos enfoques es el trabajo de Papakyriakopoulos y Mboya (2023), en el que se desarrolló un método *sociocomputacional* para el análisis de sesgos en el buscador de imágenes de Google. Para lograrlo, combinaron el análisis técnico del sistema con el uso de teorías críticas sobre el poder.

Locus

Cuando se considera el lugar en el que se impulsan las evaluaciones de algoritmos, pueden identificarse dos categorías: **interna** y **externa**. En el primer caso, el proceso se desarrolla en el marco de una organización concreta y, por lo general, lo ejecuta personal interno. Por su parte, para las evaluaciones externas no se necesita la participación o autorización de la organización que es objeto de análisis. En este caso, el proceso se desarrolla en otros entornos (universidades, medios de comunicación, etc.) de una manera independiente.

Uno de los ejemplos más conocidos, que podríamos incluir en esta última categoría, es la evaluación del sistema COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) que desarrolló el medio independiente ProPublica en el 2016. Tras el análisis de datos, se encontraron sesgos de raza en las predicciones sobre la posibilidad de reincidir en comportamientos delictivos.

Actores promotores

En relación estrecha con la dimensión anterior, en esta otra se identifican tres tipos de evaluación en función de los actores que impulsan el proceso (Costanza-Chock et al. 2022; Meßmer y Degeling 2023; Metcalf et al. 2021). Es posible que la auditoría o la evaluación de impacto sean iniciativa de **actores primarios** (***first-party***), es decir, personal de la propia organización (lo que coincidiría con el locus interno). Por lo general, los resultados de estas evaluaciones no se hacen públicos (Costanza-Chock et al. 2022). Un ejemplo es el proceso que desarrolló Amazon internamente para evaluar el algoritmo usado en la selección de personal, en el que se identificaron sesgos de género (Dastin 2018).

Por otro lado, las organizaciones pueden contratar a **actores secundarios**, es decir, a organizaciones externas (generalmente, consultoras, fundaciones, etc.) para desarrollar los procesos de evaluación (***second-party***). En este caso, los/as profesionales que llevan a cabo las evaluaciones no son completamente independientes, porque de algún modo cumplen indicaciones de las organizaciones que las contratan. Un ejemplo de este tipo de actores es O'Neil Risk Consulting & Algorithmic Auditing, la empresa fundada por la científica Cathy O'Neil para evaluar algoritmos. Entre ellos se encuentra HireVue, que se utiliza para apoyar la contratación de personal.

Finalmente, las evaluaciones ejecutadas por **terceros** (***third-party***) se caracterizan por su completa independencia respecto de la organización evaluada. Es decir, organizaciones de supervisión independientes, investigadores/as de centros académicos o periodistas evalúan sistemas algorítmicos y publican sus análisis con el fin de sensibilizar a la sociedad sobre el tema. También se han impulsado estrategias para que el público general, sin conocimientos técnicos, pueda desarrollar evaluaciones de los algoritmos que les afecten. Por ejemplo, la herramienta IndieLabel fue diseñada para que las personas usuarias puedan entrenar un modelo e identificar fácilmente si hay toxicidad en los comentarios de las plataformas de contenido (Lam et al. 2023).

Rol de actores externos

Las evaluaciones de algoritmos se pueden entender en función de la diferente implicación (rol) de las comunidades afectadas y el público general. Por un lado, las evaluaciones **no participativas** son desarrolladas por especialistas, sean externos o internos, sin considerar a otros actores externos al proceso de evaluación, que podrían aportar su visión sobre el sistema algorítmico.

Ahora bien, cada vez más personas expertas destacan la importancia de impulsar un enfoque **participativo**, como una forma de aportar diversidad a las evaluaciones, aumentar la confianza en estos procesos y contribuir a una verdadera rendición de cuentas (Groves 2022). En este caso, se trata principalmente de consultar e involucrar a las personas afectadas por las decisiones algorítmicas, organizaciones de la sociedad civil, entidades con intereses específicos y usuarios/as para detectar posibles efectos no deseados e impactos concretos no aceptables, así como para ampliar la mirada evaluadora, sobre todo, aunque no exclusivamente, entre las personas potencialmente más afectadas.

Momento

El momento de realización de las evaluaciones de algoritmos es otra de las dimensiones clave de este tipo de procesos. En estudios anteriores se ha destacado que es fundamental hacer evaluaciones durante todo el ciclo de vida de los sistemas de IA (Mökander et al. 2022; Novelli et al. 2023; Sandu et al. 2022). Considerando lo anterior, aquí se identifican dos momentos: *ex ante* y *ex post*.

Las evaluaciones algorítmicas ***ex ante*** se desarrollan en un periodo previo a la implementación del algoritmo. En ellas se deben incluir preguntas clave sobre los supuestos que están detrás de su diseño y los potenciales riesgos que existen (Ada Lovelace Institute 2020; Sloane 2021).

Una vez que ya está en marcha el algoritmo, se deben hacer evaluaciones ***ex post*** para entender los impactos reales del uso del algoritmo (Ada Lovelace Institute 2020; Eticas Consulting s.f.), incluso aquellos que no se habían previsto. Dentro de esta categoría entran aquellas evaluaciones que se desarrollan en tiempo elegido (tiempo real), es decir, durante la misma implementación del sistema algorítmico.

Orientación hacia la normativa

Las evaluaciones de algoritmos también se pueden clasificar de acuerdo con su grado de obligatoriedad u orientación hacia la normativa. En este sentido, el proceso se alinea con las inspecciones regulatorias, que se explican más adelante, en el apartado sobre metodología. Construyendo sobre la base de Kelly-Lyth y Thomas (2023), que interpretan a Burr y Leslie (2022), se proponen tres categorías en función de cuán vinculante es el cumplimiento de la normativa para las organizaciones involucradas.

La primera categoría, **obligaciones de ley**, se refiere a aquellos procesos de evaluación algorítmica que se desarrollan para cumplir con lo que se establece en la normativa de un contexto geográfico determinado. Por ejemplo, se podría desarrollar una evaluación para garantizar que se cumpla el artículo 22 del Reglamento General de Protección de Datos, que señala que toda persona interesada tiene derecho "a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles".

En segundo lugar, se encuentran los procesos de evaluación y auditoría para cumplir un **marco normativo más amplio** y menos vinculante, por ejemplo, documentos o acuerdos con principios o recomendaciones generales en torno a la IA, como la Carta europea sobre el uso ético de la inteligencia artificial en los sistemas judiciales, o la Recomendación sobre la ética de la inteligencia artificial de la Unesco, entre otros muchos. En otros casos, lo que se pretende es analizar de forma voluntaria si el uso de algoritmos se alinea con una serie de **buenas prácticas**, como la contribución a la igualdad de género o étnica o el cumplimiento de los derechos humanos. Finalmente, vale la pena mencionar la alternativa de las **certificaciones**, en cuyo caso se ofrece un sello de calidad a aquellas organizaciones que cumplen con determinados estándares éticos en el diseño e implementación de algoritmos (De Manuel et al. 2023).

Tema

Es posible que, en determinadas circunstancias, se evalúen los algoritmos en el marco de un tema muy concreto. Si bien existe una gran variedad de temas, diferentes documentos destacan de forma general los siguientes tres: uso de los datos, ética y derechos humanos, y gobernanza, entre otros posibles. Es decir, puede ser de interés mantener el foco en los **datos** que se utilizan en los sistemas algorítmicos, específicamente en términos de privacidad, transparencia y protección de datos personales[8]. También podría plantearse una evaluación de carácter más general, sobre aspectos relacionados con la **gobernanza** de los sistemas algorítmicos, sin centrarse en un aspecto específico. Y otra posibilidad es desarrollar un proceso de este tipo para evaluar **la ética aplicada y el cumplimiento de los derechos humanos**, considerando una lista de indicadores bien definidos.

En este último caso, es importante adaptarse a cada contexto. Wasilow y Thorpe (2019), por ejemplo, proponen un marco de evaluación ética para los sistemas de IA y robótica en el ámbito militar en Canadá, que incluye aspectos como el cumplimiento de códigos éticos y normativas específicas del país, así como consideraciones sobre salud, seguridad, igualdad, confianza, dignidad humana y otros. Otro ejemplo interesante es el Human Rights, Ethical and Social Impact Assessment-HRESIA (Mantelero 2018), una herramienta que combina un cuestionario de autoevaluación y la perspectiva de un comité de especialistas (cuando sea necesario) para analizar tanto los aspectos éticos como de derechos humanos en sistemas de IA.

Ámbito

Las evaluaciones de algoritmos pueden tener también un alcance diferente, dependiendo de sus objetivos, intereses y recursos. Pueden encontrarse, por un lado, evaluaciones que se centren en un **aspecto concreto del algoritmo**, como los datos de entrenamiento, el modelo subyacente o los resultados esperados, entre otras cosas (Garde Roca 2023). También pueden desarrollarse evaluaciones que abarquen **todo el ciclo de vida del sistema** y otros aspectos más amplios del contexto en el que se despliega.

[8] La Agencia Española de Protección de Datos publicó en el 2021 un documento titulado Requisitos para Auditorías de Tratamientos que incluyan IA, en el que se identifican una serie de controles de la protección de datos en tratamientos que usan componentes de IA. Disponible en: https://www.aepd.es/documento/requisitos-auditorias-tratamientos-incluyan-ia.pdf.

Nivel de acceso

Dependiendo de los actores involucrados, el locus y otros factores, se podrá tener un acceso mayor o menor a los datos que se necesitan para desarrollar evaluaciones de algoritmos. Un trabajo de Koshiyama et al. (2021) explica que existen **siete niveles de acceso**: en un extremo se encuentran los procesos ***white-box*** o 'de caja blanca' (el número 7 de la escala), en los que es posible obtener todos los detalles del sistema, mientras que en el otro extremo están las evaluaciones ***black-box*** o 'de caja negra' (el número 1 de la escala), en las que "solo se pueden hacer observaciones indirectas de un sistema" (Koshiyama et al. 2021). También existe una **zona intermedia**, ya que se da un descenso progresivo en el nivel de acceso a medida que se va pasando del número 7 al 1.

Es importante aclarar que esta clasificación se refiere al proceso de evaluación algorítmica de manera específica y no al uso del sistema algorítmico. Es decir, puede que la persona o el equipo responsable de la evaluación tenga acceso completo a un sistema algorítmico considerado *black-box* (porque sus datos y funcionamiento no están abiertos al público), en cuyo caso, la evaluación sería *white-box*. Aquí, el algoritmo no es de acceso público, pero quien desarrolla la evaluación sí tiene un acceso privilegiado al sistema y puede evaluarlo de forma completa.

Metodología

Las evaluaciones de algoritmos se llevan a cabo siguiendo diferentes métodos. Como se ha señalado previamente, el término *evaluación* alude de forma general al análisis del uso de algoritmos. Pero, al centrarse en las metodologías específicas, es importante hacer algunas distinciones conceptuales. Si bien muchas veces se usan los términos *auditoría* y *evaluación de impacto algorítmico* como sinónimos, se pueden entender de forma diferenciada, considerando las características específicas que subyacen en la metodología aplicada.

Uno de los informes más relevantes sobre esta materia, elaborado por el Ada Lovelace Institute (2020), destaca que las **auditorías** se centran en el análisis del funcionamiento de un algoritmo en relación con unos criterios específicos: por ejemplo, hipótesis concretas sobre sesgos (auditorías de sesgos o *bias audits* en inglés) o estándares que se establecen en las regulaciones (inspecciones regulatorias o *regulatory inspections* en inglés). En este caso, el proceso se desarrolla después de la implementación del sistema algorítmico, cuando ya pueden identificarse sus efectos en contextos específicos.

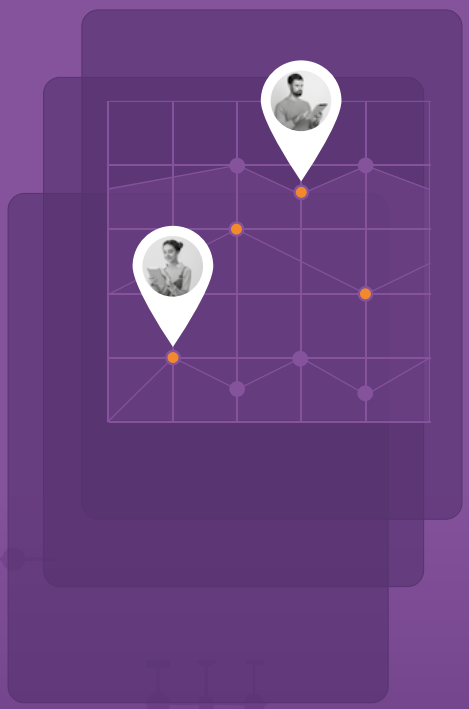
Por su parte, las **evaluaciones de impacto algorítmico** tienen un enfoque más amplio, y pueden medir los riesgos que entraña un sistema entre determinados colectivos de personas, antes o durante la implementación (análisis de riesgos o *algorithmic risk assessment*) o los impactos que se han generado después de su implementación (análisis de impacto o *algorithmic impact evaluation*).

Tabla 2.
Metodologías específicas para la evaluación de algoritmos

Auditorías		Evaluaciones de impacto	
Auditorías de sesgos	Inspecciones regulatorias	Análisis de riesgos	Análisis de impacto
Análisis de hipótesis concretas sobre sesgos que pueden existir en los datos o modelos	Análisis del cumplimiento de los estándares que se incluyen en la regulación	Medición de los riesgos de afectar a determinados colectivos (el proceso se hace antes o en las primeras fases de implementación)	Medición del impacto del algoritmo después de su implementación

Fuente de los datos: Ada Lovelace Institute

3. Métodos para el desarrollo de evaluaciones de algoritmos



Considerando la diversidad de enfoques que existen a la hora de desarrollar evaluaciones de algoritmos, se ha avanzado también en el diseño y aplicación de una gran variedad de métodos que permiten atender las particularidades de este tipo de procesos. Cada uno de ellos ofrece ventajas y desventajas, y la idoneidad de su uso dependerá de los objetivos propuestos y los recursos disponibles. En esta sección se realiza una aproximación al conjunto de métodos existentes para la evaluación de algoritmos, y se da un paso más allá al conectarlos con la tipología de dimensiones de la sección previa. Con ello se completa un mapa general sobre el estado de la cuestión, que tiene como objetivo ofrecer nuevo conocimiento y guiar los siguientes pasos para mejorar los procesos de análisis de algoritmos.

En la literatura académica y los documentos oficiales e informes de varias organizaciones se pueden identificar varios tipos de métodos para la evaluación de algoritmos. Entre las alternativas concretas más recurrentes se encuentran los siguientes: auditorías de código, *scraping*, *sock puppet*, *carrier puppet*, auditorías colaborativas, análisis estadísticos, *checklists*, encuestas a usuarios/as, *workshops* o grupos focales y estudios de caso o historias de desarrollo.

Para aportar claridad al debate, se ha elaborado una matriz que permite entender con más detalle estos métodos (véase la Tabla 3). Como se puede observar, la mayoría de los métodos tienen un carácter eminentemente técnico y poco participativo, lo que abre la puerta a ampliar la mirada hacia otras estrategias más cualitativas y holísticas en el futuro (Bandy 2021; Costanza-Chock et al. 2022). Con ello se pone el acento en la necesidad de seguir ampliando las fronteras del conocimiento en una temática que requiere un enfoque abiertamente sociotécnico, que acompase estas dos vertientes de un fenómeno que cuenta con un carácter poliédrico.

Tabla 3.
Métodos para evaluar algoritmos y relación con sus dimensiones identificadas

Dimensión	Auditorías de código	*Scraping*	*Sock puppet*	*Carrier puppet*	Auditorías colaborativas
Foco	Técnica	Técnica	Técnica	Técnica	Técnica
Locus	Interna (más probable) o externa	Interna o externa (más probable)	Interna o externa (más probable)	Interna o externa	Interna o externa
Actores promotores	Primarios (más probable), secundarios (más probable), o terceros	Secundarios o terceros	Primarios, secundarios o terceros (más probable)	Primarios, secundarios o terceros (más probable)	Primarios, secundarios o terceros
Rol de actores externos	No participativa	No participativa	No participativa	No participativa	Entre participativa y no participativa
Momento	Ex ante o ex post	Ex post	Ex ante o ex post	Ex ante o ex post	Ex ante o ex post
Orientación hacia la normativa	Obligación de ley, cumplimiento de marco normativo o buenas prácticas	Buenas prácticas	Obligación de ley, cumplimiento de marco normativo o buenas prácticas	Obligación de ley, cumplimiento de marco normativo o buenas prácticas	Obligación de ley, cumplimiento de marco normativo o buenas prácticas
Temática	Ética, DD. HH. o gobernanza	Ética, DD. HH. o gobernanza	Ética, DD. HH. o gobernanza	Ética, DD. HH. o gobernanza	Ética, DD. HH. o gobernanza
Ámbito	Aspecto concreto	Aspecto concreto	Aspecto concreto	Aspecto concreto	Aspecto concreto
Nivel de acceso	White-box	Black-box	Intermedio o black-box	Intermedio o black-box	Intermedio o black-box
Metodología	Auditoría	Auditoría	Auditoría	Auditoría	Auditoría

Dimensión	Análisis estadísticos	*Checklists*	Encuestas a usuarios/as	*Workshops* o grupos focales	Estudios de caso o historias de desarrollo
Foco	Técnica	Técnica y/o holística	Holística	Holística	Holística
Locus	Interna o externa	Interna o externa	Externa	Interna o externa	Interna o externa
Actores promotores	Primarios, secundarios o terceros	Primarios (más probable), secundarios (más probable) o terceros	Primarios, secundarios o terceros (más probable)	Primarios, secundarios o terceros	Primarios, secundarios (más probable) o terceros (más probable)
Rol de actores externos	No participativa	No participativa	Participativa	Participativa	Participativa o no participativa
Momento	Ex ante o ex post	Ex ante o ex post	Ex post	Ex ante o ex post	Ex post
Orientación hacia la normativa	Obligación de ley, cumplimiento de marco normativo o buenas prácticas	Obligación de ley, cumplimiento de marco normativo o buenas prácticas, certificaciones	Buenas prácticas	Buenas prácticas	Buenas prácticas
Temática	Uso de datos, ética, DD. HH. o gobernanza	Uso de datos, ética, DD.HH. o gobernanza	Ética, DD. HH. o gobernanza	Ética, DD. HH. o gobernanza	Ética, DD. HH. o gobernanza
Ámbito	Aspecto concreto	Sistema completo	Sistema completo	Sistema completo	Sistema completo
Nivel de acceso	Intermedio	White-box, intermedio o black-box	White-box, intermedio o black-box (más probable)	White-box, intermedio o black-box	White-box, intermedio o black-box
Metodología	Auditoría	Auditoría o evaluación de impacto	Evaluación de impacto	Evaluación de impacto	Evaluación de impacto

Fuente de los datos: elaboración propia a partir de Koshiyama et al. 2021; Pappu et al. 2021; Sandvig et al. 2014; Hamilton 2021; Raji y Buolamwini 2019; Mantelero 2018; Oswald et al. 2018; Wasilow y Thorpe 2019; Bandy 2021; DeVos et al. 2022; Dodge et al. 2021, y entrevistas a personas expertas

Auditorías de código

Las auditorías de código buscan avanzar hacia una mayor transparencia algorítmica (Sandvig et al. 2014). En este método, se analiza de manera específica el código fuente para verificar si existe algún aspecto problemático que se traduzca en resultados indeseados (por ejemplo, discriminación de determinadas poblaciones, privacidad, etc.). Se trata de un tipo de auditoría eminentemente técnica, que se puede desarrollar con más probabilidad de forma interna, principalmente por iniciativa de la organización. Esto es así porque, como se destaca en la literatura (Koshiyama et al. 2021; Sandvig et al. 2014), se necesita un acceso completo a información que puede ser sensible (*white-box*), por lo que es poco probable que las organizaciones la hagan pública.

Es posible, en todo caso, que si se contrata a una organización externa (auditorías *second-party*), se pueda compartir con ella la información para desarrollar el proceso de auditoría de forma completa. Por otro lado, al tratarse de una auditoría técnica, que se centra en un aspecto muy específico y sigue unos parámetros preestablecidos, no se contempla la participación de actores externos en el proceso. Se pueden desarrollar tanto antes de la implementación del algoritmo como después, con el objetivo de cumplir obligaciones de ley (en caso de que se especifique de esta forma en la normativa) o bien de forma voluntaria para seguir estándares y buenas prácticas.

Casos - Auditorías de código

- Una de las herramientas más conocidas para auditar modelos de aprendizaje automático es AI Fairness 360 (https://ai-fairness-360.org/), un software de código abierto, desarrollado inicialmente por IBM y actualmente bajo la iniciativa de The Linux Foundation, que puede detectar sesgos en varios elementos del ciclo de la IA, incluyendo el propio algoritmo. Otra herramienta conocida que también permite la identificación de sesgos algorítmicos es Aequitas, desarrollada por investigadores/as de la Universidad de Chicago (Saleiro et al. 2019).

- En el 2020, la empresa Pymetrics contrató a un equipo de la Northeastern University, en Estados Unidos (Wilson et al. 2020), para desarrollar una auditoría de su herramienta algorítmica de evaluación de candidaturas a empleos. El grupo de investigadores/as tuvo acceso al código fuente y a documentación adicional, y encontró que no se producían resultados discriminatorios, de acuerdo con unos parámetros muy específicos. Este caso no ha estado exento de críticas por las limitaciones de las definiciones de partida que se usaron en la auditoría para medir la discriminación (Schellmann 2021).

Scraping

En este tipo de evaluación se busca interactuar con el algoritmo de una forma intensiva para evaluar su funcionamiento y resultados. Por ejemplo, se pueden usar API (es decir, interfaces de programación de aplicaciones, que permiten la comunicación de diferentes tipos de software para obtener la información deseada) o se pueden hacer solicitudes de forma manual (por ejemplo, para evaluar el algoritmo de un buscador), pero de una manera distinta a como lo hacen personas usuarias convencionales (Pappu et al. 2021; Sandvig et al. 2014). Se trata de una evaluación de carácter técnico, pero, a diferencia de la anterior, no se obtiene acceso completo a la información sobre el código fuente, por lo que se puede desarrollar de forma externa, tanto por organizaciones contratadas como por consultoras independientes.

Generalmente, se aplica en momentos posteriores a la puesta en marcha del algoritmo y sin la participación de comunidades o personas afectadas. Es poco probable que estos procesos se exijan de manera legal, por lo que podrían entenderse, más bien, como una forma de impulsar buenas prácticas en términos de ética, derechos humanos o cualquier otro tema de interés.

Casos - *Scraping*

- Papakyriakopoulos y Mboya (2023) desarrollaron un ingenioso marco *sociocomputacional*, en el que se usan métodos de *scraping* para evaluar los sesgos y estereotipos de género y raza en el buscador de imágenes de Google. Para lograrlo, entrenaron un programa con imágenes etiquetadas, que iba enviando de forma automatizada las consultas al buscador de Google y extrayendo las respuestas del algoritmo. Luego se utilizaron métodos computacionales en combinación con un análisis cualitativo para interpretar los resultados.

- En una investigación del 2018, Kulshrestha et al. (2019) analizaron sesgos políticos en los resultados de búsqueda de Twitter y Google, en relación con las primarias presidenciales de Estados Unidos en el 2016. Para conseguirlo, interactuaron directamente con las plataformas y obtuvieron los datos que estaban disponibles de forma pública y abierta.

Sock puppet

En este método, se utilizan programas para hacer las veces de usuarios/as de un sistema y evaluar las decisiones que se toman en función de sus perfiles. A diferencia del *scraping*, las evaluaciones *sock puppet* permiten obtener información más detallada sobre las variables específicas que se quieren estudiar (Pappu et al. 2021), si bien existe debate en torno a la ética de este tipo de métodos (Sandvig et al. 2014).

Es posible que alguna organización esté interesada en desarrollar de forma interna este tipo de evaluación, para detectar posibles problemas en el algoritmo antes o después de su implementación, y cumplir con alguna ley específica o fomentar buenas prácticas. Sin embargo, es más probable que ocurra a petición de una organización contratada o por iniciativa de investigadores/as independientes. Es una alternativa para evaluar un algoritmo cuando no se cuenta con información detallada sobre el código fuente (por eso, puede considerarse como *black-box* o en un estadio intermedio del espectro de acceso), y no es necesaria la participación de comunidades afectadas en el proceso.

Casos - *Sock puppet*

- Las investigadoras Eriksson y Johansson (2017) crearon 288 cuentas de Spotify (bots), la mitad de ellas registradas como hombres y la otra mitad como mujeres. Su intención era verificar si existían sesgos de género en las recomendaciones musicales de la plataforma.

- Recientemente, un grupo de investigadores/as (Srba et al. 2023) utilizaron la metodología de evaluaciones *sock puppet* para estudiar los riesgos de caer en un filtro burbuja de desinformación en YouTube. En este caso, se programaron bots para ponerse en el lugar de personas usuarias de la plataforma, y se analizaron las búsquedas, los resultados de la página principal y las recomendaciones de vídeos.

Carrier puppet

Este método es parecido al anterior, aunque, en lugar de ponerse en el lugar de un/a usuario/a final, se utilizan programas para hacer las veces del desarrollador/a. Es decir, se hacen pruebas para detectar posibles problemas en un estadio intermedio del desarrollo del sistema, no con el producto final (Raji y Buolamwini 2019). En cuanto a sus características y posibilidades, son muy parecidas a las del método anterior.

Quizás una organización pueda estar interesada en ejecutar este tipo de evaluación antes de lanzar un producto al público, pero es mucho más probable que se realice a petición de actores externos, para impulsar buenas prácticas y sensibilizar en torno a temas relacionados con ética, derechos humanos y otros. Tampoco se contempla la participación de comunidades afectadas directamente ni se necesita obtener la información completa sobre el código, si bien es necesario cierto grado de acceso para desarrollar el proceso de forma adecuada.

Caso - *Carrier puppet*

- El caso más conocido es la investigación denominada *Gender Shades* (Buolamwini y Gebru 2018), en la que se utilizó el método *carrier puppet* para detectar sesgos de género y raza en sistemas de reconocimiento facial. En este caso, las investigadoras encontraron que estos sistemas son menos acertados en la identificación de mujeres negras, lo que podría traducirse en efectos negativos para esta población.

Auditoría colaborativa

La auditoría colaborativa funciona de forma parecida a la de *sock puppet*, con la diferencia de que se contrata a personas usuarias para hacer las pruebas con el sistema (Sandvig et al. 2014). Este tipo de método puede usarse tanto de forma interna como externa, sea por iniciativa de la propia organización (con personal interno o externo) o por interés de personal investigador independiente. Cuando el planteamiento es interno, se realiza antes de la implementación para verificar cualquier comportamiento problemático, o bien se ejecuta después de la puesta en marcha, especialmente en caso de que se realice por parte de una organización evaluadora externa. Si se desarrolla por iniciativa de la propia organización, puede ser una buena estrategia para cumplir con obligaciones legales o marcos normativos, al tiempo que se impulsan las buenas prácticas.

En cuanto a la participación, hay que hacer algunas clarificaciones. Se cataloga como un estadio intermedio entre participativa y no participativa porque, si bien se incluye a usuarios/as de los algoritmos, por lo general son diseños experimentales en los que cada grupo debe seguir unas instrucciones concretas. Es decir, se toman en cuenta las experiencias de personas usuarias, pero no necesariamente se incluye a personas afectadas o potencialmente afectadas realmente por el algoritmo, a menos que así lo establezca el diseño y enfoque de la auditoría.

Casos - Auditoría colaborativa

- En el estudio de Spyridou et al. (2022), participaron 18 personas, divididas en dos grupos. Cada uno/a instaló un *plug-in* en su buscador e interactuó con el portal MyNews de acuerdo con unas instrucciones específicas. La información recabada permitió analizar el comportamiento de los algoritmos de recomendación de noticias.

- El portal de periodismo independiente The Markup desarrolló el proyecto Citizen Browser para auditar algoritmos de redes sociales, en concreto, de Facebook. A cambio de una compensación económica, un total de 1.000 residentes de Estados Unidos instalaron en sus ordenadores personales una aplicación que permitía la recolección de información sobre su uso de la red social. Con los datos recopilados, se han publicado investigaciones sobre una variedad de temas como el aborto, estafas relacionadas con criptomonedas, o el contenido sobre la extrema derecha, entre otros (https://themarkup.org/series/citizen-browser).

Análisis estadístico

Otro método que se usa con frecuencia en los procesos de evaluación de algoritmos es el análisis estadístico de datos y resultados del sistema. Este tipo de procesos se pueden desarrollar tanto a nivel interno como externo, bajo la iniciativa de las propias organizaciones, consultoras externas o evaluadores/as independientes. En este caso, es necesario el acceso a determinados datos, por lo que puede enmarcarse en un nivel intermedio (entre *white-box* y *black-box*). Por ejemplo, si una organización dispone de datos relevantes, es factible que personal investigador independiente use estos datos para hacer análisis estadísticos sobre determinadas variables de interés (Hamilton 2021). Pero también es posible que las propias organizaciones decidan usar los datos para realizar este tipo de estudios de manera interna, antes o después de poner en funcionamiento un algoritmo.

Este proceso permite hacer inferencias sobre un aspecto concreto del sistema (sesgos, discriminación, privacidad, etc.), pero es más limitado que el análisis del código o el modelo. Al igual que en otros métodos, tampoco se necesita la participación de comunidades externas, y puede usarse como complemento de otros métodos con el fin de cumplir obligaciones normativas o impulsar de forma voluntaria el desarrollo ético de la IA en la organización.

Casos - Análisis estadístico

- El análisis estadístico se incluye como parte del Model Risk Audit, propuesto por Munz et al. (2023) para evaluar modelos de IA en cuatro categorías: 1) robustez, 2) seguridad y privacidad, 3) explicabilidad y sesgos, y 4) rendimiento e integridad metodológica. Es especialmente relevante para el caso de la explicabilidad y los sesgos.

- La profesora Melissa Hamilton, de la Universidad de Surrey, en el Reino Unido, desarrolló un estudio sobre Public Safety Assessment (Hamilton 2021), una herramienta algorítmica que se usa en Estados Unidos para realizar predicciones en el contexto de investigaciones preliminares (antes de ir a juicio). Uno de los aspectos que intenta predecir este sistema es el riesgo de que una persona que esté bajo investigación no se presente a las citas en el juzgado, antes de que se haya tomado una decisión definitiva sobre su caso. Para hacer la evaluación del algoritmo, la investigadora analizó estadísticamente las predicciones de la herramienta y los datos reales de los casos en tres estados de Estados Unidos.

Checklists

Las listas de verificación o *checklists* permiten recoger información relevante sobre el uso de algoritmos en relación con una serie de indicadores predefinidos. Pueden incluir preguntas abiertas o cerradas (por lo general, con respuesta sí/no), o también pueden construirse tablas para recopilar información específica. En este caso, existe una amplia variedad de alternativas. Estos instrumentos se utilizan de forma interna, como una manera de hacer comprobaciones antes de lanzar un algoritmo, o bien se le encarga a una consultora externa el desarrollo del proceso.

Por su parte, investigadores/as independientes u organismos supervisores pueden usar este tipo de listas para recabar la información necesaria para evaluar un algoritmo, siempre que se cuente con la colaboración de la organización. En algunos casos, por ejemplo, se puede requerir la realización de entrevistas a desarrolladores/as, personal de la organización, etc., o el uso de herramientas técnicas para analizar detalles del algoritmo. Dependiendo de cómo se diseñe, puede permitir el estudio del sistema completo y su relación con el contexto en el que se despliega, con el foco puesto en algún tema concreto (uso de datos, derechos humanos o medioambiente, por ejemplo) o con preguntas más generales.

Casos - *Checklists*

Existen varias herramientas para desarrollar listas de verificación en materia de supervisión de algoritmos. Se enumeran a continuación algunos de ellos:

- Modelo *algo-care* (Oswald et al. 2018)
- Human Rights, Ethical and Social Impact Assessment-HRESIA (Mantelero 2018)
- Ethics Assessment Framework (Wasilow y Thorpe 2019)
- After-Action Review for AI (Dodge et al. 2021)
- Herramienta de evaluación de impacto algorítmico del Gobierno de Canadá (https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html)
- Herramienta de evaluación de impacto algorítmico del Chief Information Officers Council de Estados Unidos (https://www.cio.gov/aia-eia-js/#/)

Encuestas a personas usuarias

Las encuestas a usuarios/as reales, también denominadas *auditorías no invasivas* (Sandvig et al. 2014), pueden ser útiles para conocer aspectos de interés sobre el impacto de un algoritmo que ya se haya implementado. En este método, la información recabada permite hacer inferencias sobre el funcionamiento del sistema a través de las percepciones de las personas usuarias, pero no se pueden establecer relaciones de causalidad entre las variables estudiadas (Ibid.). Es decir, las encuestas se pueden usar como complemento de otros procedimientos más técnicos, que permitan obtener información sobre el funcionamiento real del sistema.

Cuando no es posible emplear otro método, las encuestas son útiles para conseguir datos que permitan entender de forma holística el impacto de un algoritmo en la vida real.
Eso quiere decir que, si bien una organización puede impulsar el desarrollo de estos procesos con el fin de conocer el impacto de un algoritmo, cualquier evaluador/a independiente tiene la oportunidad de desarrollar este tipo de estudios para sensibilizar sobre la materia.
Se cuenta, además, con la libertad suficiente para plantear cuestiones sobre temas diversos, en torno a la ética, los derechos humanos, el medioambiente y otros.

Caso - Encuestas a personas usuarias

- Un equipo de investigadores/as de las universidades de Stanford y Pensilvania y del Instituto Tecnológico de Georgia (Lam et al. 2023) diseñaron la plataforma Intervenr para hacer evaluaciones sociotécnicas en buscadores de Internet. Como parte de su investigación, complementaron la observación del comportamiento de los usuarios/as (lo que equivaldría a una auditoría colaborativa) con encuestas sobre su experiencia.

Workshops o grupos focales

En el mundo académico se está empezando a abogar por la incorporación de usuarios/as reales en procesos de evaluación algorítmica, empleando métodos propios de la investigación cualitativa (DeVos et al. 2022; Groves 2022). Además de las entrevistas, una de las alternativas que se propone es la organización de talleres o *workshops*, en los que las personas participantes puedan expresar sus ideas y experiencias en torno al impacto de los algoritmos.

En este método, hay también varias posibilidades. Por ejemplo, una organización podría coordinar un encuentro de este tipo para que personas potencialmente afectadas prueben libremente un sistema y discutan sus opiniones antes de la implementación. O también una consultora o investigador/a independiente puede desarrollar uno de estos *workshops* para evaluar un algoritmo que ya está puesto en marcha. Este tipo de actividades puede aportar un enfoque más holístico a la evaluación y explorar una variedad de temas desde una perspectiva diversa.

Casos - *Workshops* o grupos focales

- La combinación de entrevistas para pensar en voz alta (recopilación de opiniones mientras se está usando un sistema algorítmico), diarios y *workshops* permitió a la investigadora DeVos y colaboradores/as (2022) obtener la perspectiva de usuarios/as reales en la evaluación de impactos algorítmicos.

- Un equipo de investigación compuesto por la consultora Eticas, la Universidad Pompeu Fabra de Barcelona y ALPHA Telefónica evaluó la app REM!X, desarrollada por Telefónica Innovación Alpha para ofrecer recomendaciones sobre bienestar (Galdon Clavell et al. 2020). Se siguieron cuatro estrategias para ejecutar la evaluación: análisis de recomendaciones del algoritmo, revisión documental, etnografía digital (un tipo de investigación que analiza las relaciones sociales que se producen en el entorno online) mediante el estudio de los mensajes de *feedback* de las personas usuarias, y cinco grupos focales en los que participaron evaluadores/as, así como ingenieros/as y desarrolladores/as de la app.

Estudios de caso o historias de desarrollo

Para atender a la complejidad de los sistemas algorítmicos, también se está proponiendo avanzar en el uso de estudios de caso o historias de desarrollo (Bandy 2021) con un enfoque asimilable al de la etnografía digital. La primera alternativa consiste en el análisis en profundidad de un caso concreto, preferiblemente a través del uso de observación directa, entrevistas y otros métodos cualitativos que permitan captar las dimensiones organizacionales y sociotécnicas de los algoritmos.

En cuanto a la segunda opción, se trata de reconstruir la historia del desarrollo de un algoritmo, para encontrar detalles que permitan entender sus problemas y potenciales soluciones. Al igual que en la opción anterior, este tipo de evaluaciones pueden hacerse por iniciativa de las propias organizaciones o por parte de actores externos, si bien es quizás más probable que las desarrollen consultoras o investigadores/as independientes. Dependiendo del enfoque, pueden contar con la participación de comunidades afectadas o bien centrarse únicamente en el personal de las organizaciones que diseñan e implementan los algoritmos. Por otra parte, para obtener toda la información necesaria, es preferible que el proceso se desarrolle después de la puesta en marcha del sistema.

Casos - Estudios de caso o historias de desarrollo

- El investigador DeVito (2017) analizó notas de prensa, patentes y documentos oficiales relacionados con Facebook para reconstruir la historia de desarrollo e identificar los principales valores en el algoritmo que podrían explicar la selección de contenido en la sección de noticias (*feed*) de Facebook.

- La consultora Eticas desarrolló una auditoría externa del sistema Viogén, utilizado en España para predecir el riesgo de que una mujer vuelva a ser víctima de violencia de género. El equipo auditor no tuvo acceso a los datos originales que se habían usado, pero se desarrollaron análisis estadísticos con datos secundarios, entrevistas a 31 mujeres y un cuestionario con siete abogados/as. Si bien no se contó con la perspectiva del personal público que usa el sistema, se pudieron reconstruir sus principales características y limitaciones a través de los testimonios de las mujeres que tuvieron experiencias con esta herramienta durante la atención de sus casos.

4. Ecosistema de evaluación de algoritmos y niveles de gobernanza de la rendición de cuentas algorítmica

Los procesos de evaluación de algoritmos no se desarrollan de manera aislada en las organizaciones, sino que forman parte de un ecosistema más amplio con varios niveles de gobernanza, en el que participan diferentes ámbitos (público, privado y social), sectores o áreas de actividad (salud, educación, transportes, banca, energía, seguridad, etc.) y actores (tanto directa como indirectamente relacionados con el proceso). Es decir, se está configurando un ecosistema de evaluación algorítmica, formado por una serie de componentes que interactúan entre sí y que, a su vez, se relacionan con el entorno más amplio de la IA y la rendición de cuentas, tanto a nivel nacional en cada país como a escala internacional (Percy et al. 2021; Stahl et al. 2023).

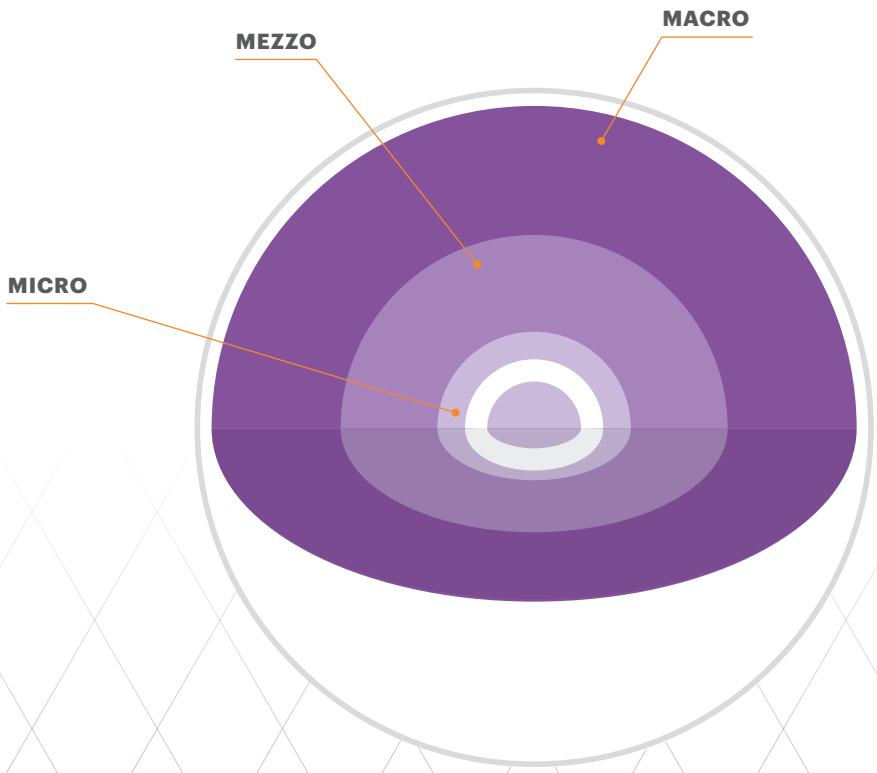
Aunque todavía queda mucho camino por delante, se ha señalado que este ecosistema se encuentra en rápido crecimiento y, por tanto, merece la atención de los reguladores, desarrolladores, empresas, consultoras, Administraciones públicas y el mundo académico, pero también de las personas usuarias de los sistemas y de la sociedad civil en su conjunto, para promover estándares específicos (Costanza-Chock et al. 2022) que contribuyan a desarrollar prácticas comúnmente aceptadas, considerando la variedad de intereses presentes.

A continuación, se ofrece una aproximación a este ecosistema emergente, específicamente, a partir de un enfoque que destaca las relaciones sociales subyacentes en estos procesos complejos de rendición de cuentas y los resultados de nuestras entrevistas y de la revisión documental realizada. Atendemos tres niveles o capas de gobernanza de la evaluación de algoritmos (*macro*, *mezzo* y *micro*), que cuentan con diferentes componentes que se desarrollan a continuación.

Una cuestión de términos, nuevamente

A partir del enfoque de Bovens (2007), se podría definir la rendición de cuentas algorítmica como la relación entre los actores que diseñan o usan algoritmos y los foros (*forums*) que han de hacer cumplir las normas de conducta de los actores participantes. Lo anterior supone la existencia de determinados requisitos de actuación y resultados por los que los actores han de rendir cuentas y por los que puede que deban responder haciendo frente a consecuencias por los usos de los algoritmos. Esas relaciones se pueden considerar de diferentes maneras, por ejemplo, según el nivel de obligación (vertical, horizontal o diagonal) o la naturaleza de los actores (individual, colectiva, jerárquica o corporativa).

Figura 1.
Los tres niveles o capas de gobernanza de la evaluación de algoritmos



Fuente de los datos: elaboración propia

Nivel *macro* de gobernanza

Ámbitos público, privado y social

En el ecosistema de las evaluaciones de algoritmos, se debe considerar la interacción entre el sector público, el sector privado y el sector social (o tercer sector). Como se destaca en algunas publicaciones académicas y documentos oficiales, es indispensable prestar atención a las relaciones de interdependencia que existen entre Estado y mercado en el desarrollo e implementación de la IA (Stahl et al. 2023). Si bien estos intercambios pueden fluir de una forma menos sistematizada o informal, se ha apuntado la necesidad de que el sector público asuma el liderazgo en estas dinámicas, puesto que el sector empresarial por sí mismo difícilmente podrá autorregular todos los riesgos e impactos inherentes a los algoritmos (Baeza-Yates y Matthews 2022).

Adicionalmente, la necesidad de incorporar a la sociedad civil en dichas dinámicas deviene por el impacto creciente de los algoritmos en cada vez más ámbitos de la vida humana, así como por la conciencia compartida de que la ciudadanía debe disponer de información y adoptar decisiones bien documentadas sobre sus relaciones con la IA.

Es importante destacar que los equilibrios en la relación entre el sector público y el sector privado dependerán del contexto específico. Como se ha apuntado en secciones previas, existen varios modelos regionales de desarrollo de la IA a nivel global (a primera vista, en Norteamérica, la Unión Europea y China) que dan cuenta de los distintos acuerdos institucionales y tradiciones económicas, políticas, sociales, culturales, etc. En algunos casos, el Estado tendrá un mayor peso y se orientará hacia el liderazgo en la elaboración de la normativa de uso y evaluación de la IA; en otros, las lógicas del mercado podrán incidir en la forma en la que se intentan regular las evaluaciones de algoritmos, a partir de una visión más centrada en la autorregulación y la generación continuada de innovaciones en el campo de la IA.

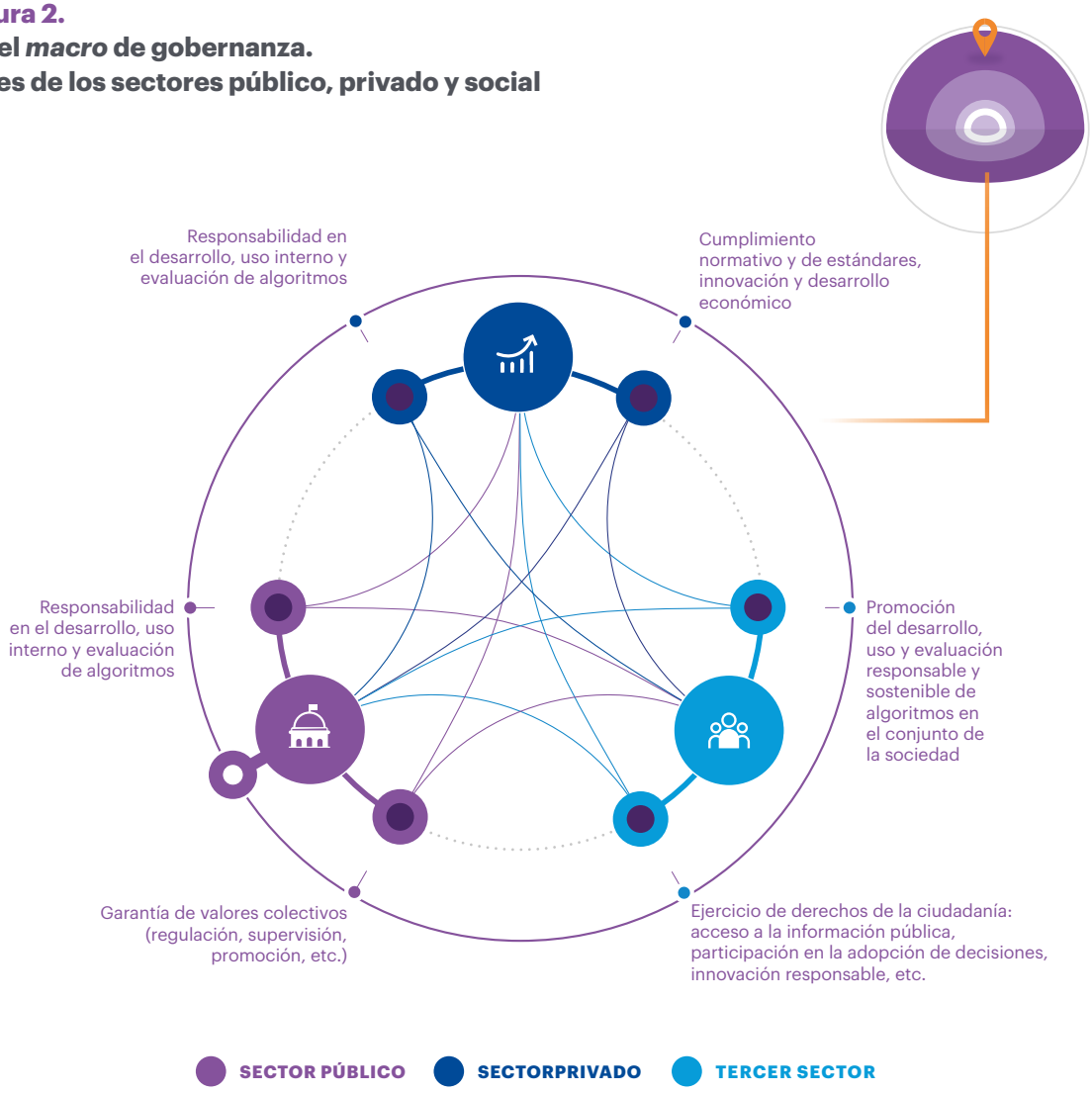
Junto con lo anterior, el papel de la sociedad civil en torno a la evaluación de la IA es importante en la medida que puede servir de punto de referencia para determinar cómo se equilibra la relación entre estados y mercados, y cuál es la posición de ciertos valores que determinan la convivencia colectiva, tales como la igualdad, la libertad, etc.

Dentro de este nivel *macro* de la gobernanza de la rendición de cuentas algorítmica, el papel del sector público es doble, como usuario de los algoritmos y como garante de los valores colectivos, independientemente de la posición del mercado y la sociedad civil en cada contexto geográfico. De entrada, ya se han puesto en marcha diferentes intervenciones públicas para mejorar la implementación de este tipo de procesos de evaluación de algoritmos. Tal y como señala un documento reciente (Basu et al. 2021), se pueden identificar hasta siete tipos de actuaciones (o políticas) públicas en relación con la rendición de cuentas algorítmica que, en algunos casos, coinciden con los métodos ya considerados en este documento. Tal es el caso de los principios y guías, las evaluaciones de impacto o la puesta en práctica de auditorías técnicas e inspecciones regulatorias.

Ahora bien, en otros casos tienen un alcance diferente, ya que, en realidad, más que métodos de evaluación, se refieren a mecanismos que han diseñado los gobiernos y las Administraciones sobre cómo enmarcar la relación con el sector privado y la sociedad civil. Esto incluye las prohibiciones y moratorias, la promoción de la transparencia pública, la existencia de órganos independientes

de control externo, el derecho a ser escuchado y apelar decisiones, o las condiciones de la contratación pública de IA. Aquí podemos encontrar una relación directa con el papel del sector privado[9] y la sociedad civil[10], de manera que es necesario diseñar mecanismos que faciliten la gobernanza del proceso de evaluación de algoritmos, integrando lo más armoniosamente posible esos tres ámbitos que intervienen (sector público, privado y sociedad civil).

Figura 2.
Nivel *macro* de gobernanza.
Roles de los sectores público, privado y social



Fuente de los datos: elaboración propia

[9] Por ejemplo, en las condiciones de la contratación pública de IA o las prohibiciones a las empresas sobre el desarrollo de determinados algoritmos de alto riesgo.

[10] Por ejemplo, el derecho a ser escuchado y apelar decisiones, como plantea la Regulación General de Protección de Datos de la UE, o la explicación accesible para el público de las decisiones basadas en algoritmos, de manera que cualquier persona pueda comprender su contenido, independientemente de sus características personales o nivel de educación formal.

Nivel *mezzo* de gobernanza

Sectores de actividad (salud, educación, transportes, comunicaciones, banca, energía, seguridad...)

Gracias a los avances acelerados de la IA, prácticamente cualquier sector de actividad es susceptible de automatizar sus procesos, en mayor o menor medida, a través del uso de algoritmos. Es necesario articular por ello mecanismos de gobernanza de la rendición de cuentas algorítmica adecuados dentro de cada uno de esos sectores de actividad intermedios (nivel *mezzo*), considerando que se pueden producir situaciones diferenciadas, si bien también es necesaria una cierta homologación entre sectores, desde la salud, educación, seguridad, etc. (generalmente, más cercanos al sector público), hasta los transportes, energía, banca, telecomunicaciones, etc. (generalmente, más cercanos al sector privado).

En este sentido, las evaluaciones algorítmicas no deben reservarse a algunos ámbitos específicos, sino que deben permear en todo el ecosistema de la IA y centrarse en efectos más amplios, considerando también los otros sectores de actividad con los que puedan estar relacionados. Desde este reconocimiento, y como destacaron algunas de las personas entrevistadas para este proyecto, es importante prestar atención especial a aquellos sectores que usan algoritmos que pueden tener un impacto más directo en la vida de las personas (por ejemplo, la salud o los servicios financieros), aunque también se deban tomar en cuenta otros criterios de interés, como las implicaciones para el medioambiente, los efectos en los derechos de determinados colectivos de personas vulnerables o los propios efectos sobre la evolución de cada sector de actividad.

La propuesta de Ley de IA de la UE ofrece una hoja de ruta para avanzar en este asunto. Si bien se incluyen estándares generales, se destacan algunos sectores de alto riesgo, como el orden público, la Administración de justicia, la migración y gestión de fronteras, y el acceso a servicios básicos. En estas áreas prioritarias, el desarrollo de evaluaciones de algoritmos se hace aún más indispensable. Si bien no existe un consenso al respecto, además de las orientaciones generales, podría ser necesario el establecimiento de estándares específicos para cada sector, de manera que puedan atenderse las particularidades de las áreas concretas[11].

Si se sigue este planteamiento, los procesos de evaluación deberán ser prioritarios cuando se usen algoritmos en actividades específicas. Por ejemplo, entre las acciones más invasivas se encuentran la vigilancia de personas (a través, por ejemplo, del reconocimiento facial, entre otros métodos), la elaboración de perfiles, la clasificación de individuos o la adopción de decisiones automatizadas sobre la asignación de derechos de acceso a servicios públicos o beneficios sociales. Algunas de estas prácticas, como la vigilancia en tiempo real a través de información biométrica, están prohibidas en la propuesta de Ley de IA de la UE, por considerarse que entrañan riesgos inaceptables. En todo caso, si se observa el ecosistema de las evaluaciones de forma holística, es indispensable que todos los sectores estén sensibilizados sobre los posibles impactos que se pueden generar con el uso de este tipo de sistemas para determinados propósitos, y tomen las medidas necesarias para la prevención y mitigación de esos problemas.

[11] Un ejemplo en esta línea es la evaluación de impacto algorítmico que desarrolló el Ada Lovelace Institute en el contexto del sistema de salud del Reino Unido (Groves 2022), que demostró la importancia de adaptar sus procesos a las características propias de cada ámbito y espacio geográfico.

Figura 3.
Nivel *mezzo* de gobernanza.
Incidencia en diferentes áreas de actividad

1
Salud
Algoritmos para el diagnóstico y seguimiento de pacientes en hospitales, así como dispositivos y robots para el acompañamiento en el hogar.

2
Educación
Sistemas para el apoyo a docentes en la evaluación del alumnado. Usos de IA generativa para el desarrollo de actividades de aprendizaje.

3
Justicia
Sistemas de evaluación de riesgos y predicciones de reincidencia.

4
Seguridad
Sistemas de reconocimiento facial para la identificación de personas que han cometido delitos. Sistemas de policía predictiva.

5
Migraciones
Sistemas de reconocimiento facial en puestos de control de fronteras. Sistemas para la evaluación automatizada de solicitudes y trámites de migración.

6
Transporte
Automóviles autónomos. Sistemas para automatizar procesos logísticos.

7
Infraestructuras
Herramientas de IA para el diseño de infraestructuras. Sistemas de mantenimiento predictivo.

8
Telecomunicaciones
Uso de IA para mejorar redes y servicios. Sistemas para la optimización y ahorro de recursos.

9
Energía
Algoritmos para el análisis de datos en tiempo real y para el mantenimiento predictivo de infraestructuras. Sistemas de prevención de catástrofes y compartición de recursos energéticos.

Fuente de los datos: elaboración

Nivel *micro* de gobernanza

Actores con un papel directo o indirecto (organizaciones implementadoras, analistas externos, desarrolladores, personas usuarias...)

Como se ha destacado en secciones anteriores, las evaluaciones de algoritmos pueden clasificarse de acuerdo con los actores (tanto organizaciones como personas) que desarrollan el proceso, pero aquí ampliamos el foco a quienes tienen alguna relación, directa o indirecta. En el primer sentido, Costanza-Chock et al. (2022) identifican tres tipos de actores: personal interno de las organizaciones usuarias (*first-party*), consultoras y otras organizaciones especializadas (*second-party*) y organismos evaluadores o personal investigador independiente (*third-party*). Estos tres tipos de actores conforman una parte clave del ecosistema de las evaluaciones algorítmicas y desempeñan roles muy diferenciados.

En el segundo sentido, existen otros actores que, si bien pueden tener un papel no directamente relacionado con el proceso evaluador en sí mismo, también se deben considerar como parte de este nivel *micro* de gobernanza, incluyendo tanto a personas usuarias (y no usuarias) de los sistemas como a otras empresas y organizaciones, organismos reguladores o supervisores, empresas desarrolladoras, así como otras organizaciones y personas de la sociedad civil involucradas.

Actores con un papel directo

Las **organizaciones que implementan algoritmos *(first-party)*** son un agente central del proceso de evaluación algorítmica y cuentan con una responsabilidad clave en la puesta a disposición de los datos sobre el funcionamiento técnico y no técnico de los sistemas algorítmicos, y las dinámicas organizativas que influyen en su despliegue. Esta información es fundamental para garantizar una verdadera rendición de cuentas y un desarrollo responsable y ético de los algoritmos. Junto con todas las empresas de los diferentes sectores de actividad, aquí habría que poner especial énfasis en las Administraciones públicas que implementan algoritmos, dado que en este caso específico se trata de organizaciones con una responsabilidad reforzada por el tipo de áreas de actividad en que se desempeñan (en muchos casos, sin capacidad de elección por parte de las personas usuarias), así como por los efectos directos de dichos algoritmos sobre la vida de las personas y las consecuencias sobre la igualdad y la libertad de la ciudadanía.

En consecuencia, las organizaciones implementadoras son las principales responsables de que el funcionamiento de los algoritmos que utilizan en sus actividades cumpla con la normativa vigente y, más allá de ello, responda a unos estándares básicos sociales, éticos y de derechos humanos que cada vez más organizaciones internacionales demandan.

Por otro lado, las **consultoras y otras organizaciones especializadas en evaluaciones algorítmicas (*second-party*)** aportan una mirada externa y los conocimientos necesarios para desarrollar este tipo de procesos de manera adecuada, bajo demanda de las organizaciones implementadoras, si bien cumpliendo unos estándares profesionales, por ejemplo, mediante certificaciones o sellos que garanticen la calidad de las evaluaciones. Sin embargo, se ha destacado que, sin los estándares apropiados y los mecanismos de regulación y control necesarios, existe un cierto riesgo de que este tipo de empresas especializadas no

tengan la libertad suficiente para desarrollar su trabajo de forma independiente y que las organizaciones usuarias perciban estos procesos como una oportunidad para limpiar su imagen, algo que podríamos denominar *algorithm washing* o *algowashing*, sobre todo, en casos dudosos sobre la existencia de sesgos o vulneraciones de la protección de los datos personales (Goodman y Trehu 2022; Schellmann 2021; personas entrevistadas).

En definitiva, estos actores son clave para llevar a cabo las labores de evaluación de algoritmos, sobre todo desde una perspectiva técnica, si bien es cierto que su previsible proliferación en el futuro requerirá algún tipo de inspección o fiscalización por parte de entidades de supervisión o estandarización, sean públicas o privadas, que permitan profesionalizar este tipo de actividad, así como armonizar su despliegue en diferentes contextos en aras de aumentar la confianza empresarial y social.

En tercer lugar, se encuentran **las organizaciones evaluadoras o el personal investigador externo (*third-party*)**, lo que incluye diferentes iniciativas independientes de supervisión de algoritmos, así como personal del sector académico, activistas, periodistas, etc. Estos actores tienen la ventaja principal de ser independientes y realizar su labor sin necesidad de que exista una demanda previa de la organización evaluada. En todos estos casos también se plantean determinadas cuestiones, por ejemplo, el hecho de que no siempre cuentan con los recursos o el acceso directo a los datos de la organización implementadora para desarrollar su labor (Costanza-Chock et al. 2022).

Adicionalmente, estos actores puede que no se sientan sometidos a códigos de conducta profesionales a la hora de desarrollar sus evaluaciones o no expliciten sus potenciales conflictos de interés. En resumen, sería deseable que este grupo de actores pueda tener un papel activo en los procesos de rendición de cuentas algorítmica, al mismo tiempo que se respeta la autonomía de las organizaciones implementadoras, por ejemplo, mediante acuerdos para la cesión de datos a cambio de los resultados de las evaluaciones, o la promoción de mediciones periódicas como oportunidad de aprendizaje conjunto.

Actores con un papel indirecto

Además de los tres tipos de actores mencionados, también emergen otros que, a primera vista, cuentan con un papel indirecto en la gobernanza de la rendición de cuentas algorítmica, destacando las **personas usuarias**. Como se ha apuntado en otro lugar (Stahl et al. 2023), cuando se lleva a cabo una evaluación algorítmica, se debe tener en cuenta cómo se interrelacionan las experiencias y percepciones de quienes utilizan los algoritmos, de manera que se pueda obtener una panorámica completa sobre sus riesgos e impactos, especialmente, entre colectivos que puedan estar especialmente expuestos a ellos.

En este sentido, muchos de los métodos que se han apuntado en secciones anteriores tienen como objetivo esencial integrar en el ecosistema de las evaluaciones algorítmicas a las personas usuarias de los sistemas, incluso a las que no lo son (por ejemplo, para llevar a cabo experimentos y testar hipótesis). Y lo anterior no es más que una garantía de que los procesos de análisis integran una dimensión social, además de otros aspectos más organizativos o técnicos, de forma que se puedan completar con actuaciones centradas en las personas, protegiendo más adecuadamente sus derechos.

En este caso, no solo se considera a usuarios/as en el ámbito externo a las organizaciones, sino también al personal de las organizaciones que implementan algoritmos y que tienen un contacto directo con estos sistemas. Por ejemplo, el personal de una organización que usa herramientas basadas en algoritmos para tomar decisiones sobre los servicios que se prestan al público. Es importante, en este caso, abordar las diversas aristas que existen en torno al uso de los algoritmos en una variedad de contextos.

Las **entidades de regulación o supervisión** algorítmica son fundamentales en el estadio inicial del proceso, pero también en el resto del ciclo de vida de los algoritmos. Como ya se ha destacado en otras partes de este trabajo, en diferentes contextos se ha puesto el acento en la necesidad de crear registros públicos de algoritmos, así como de promover la puesta en marcha de agencias de regulación o supervisión algorítmica que contribuyan, entre otras cosas, a los procesos de evaluación.

En este sentido, uno de los casos recientes más avanzado consiste en la previsión de creación de la Agencia Española de Supervisión de la Inteligencia Artificial[12], la primera de esta naturaleza en el contexto de la UE, si bien se pueden identificar otras iniciativas que estarían en una línea semejante en Canadá[13], Reino Unido[14] o Singapur[15]. En resumen, las agencias públicas de supervisión tendrán un papel esencial en el ecosistema de rendición de cuentas algorítmica, a través de métodos directos o indirectos de actuación, lo que dependerá del contexto institucional y regulatorio concreto, así como de las prioridades de cada país en relación con esta materia.

Las **empresas desarrolladoras de algoritmos** también tienen un papel crítico a la hora de entender las dinámicas de evaluación algorítmica. Estas organizaciones se encuentran en el origen del proceso, dado que son las que se encargan de la creación de los algoritmos que implementarán en un momento posterior otras empresas y las Administraciones públicas.
Al margen de las asociaciones profesionales que se puedan crear en el futuro, resulta evidente que estas empresas cuentan con una responsabilidad en el proceso de evaluación de algoritmos, ya que deben cumplir con los principios, estándares éticos y regulaciones técnicas que progresivamente van proliferando en diferentes contextos institucionales.
En definitiva, se requiere que este tipo de actores tenga muy presente desde el principio que su actividad estará especialmente supervisada, al mismo tiempo que se defienden sus derechos de propiedad sobre los algoritmos que generen o se promueve su capacidad para desarrollar futuras innovaciones en este campo de actividad.

[12] Se trata de una entidad cuyo estatuto se aprobó a través del Real Decreto 729/2023, de 22 de agosto. En el artículo 4 de su anexo se indica que el organismo "tendrá la función de inspección, comprobación, sanción y demás que le atribuya la normativa europea que le resulte de aplicación y, en especial, en materia de IA".

[13] Office of the Chief Information Officer (OCIO), Treasury Board of Canada Secretariat (TBS) con su Algorithmic Impact Assessment tool: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

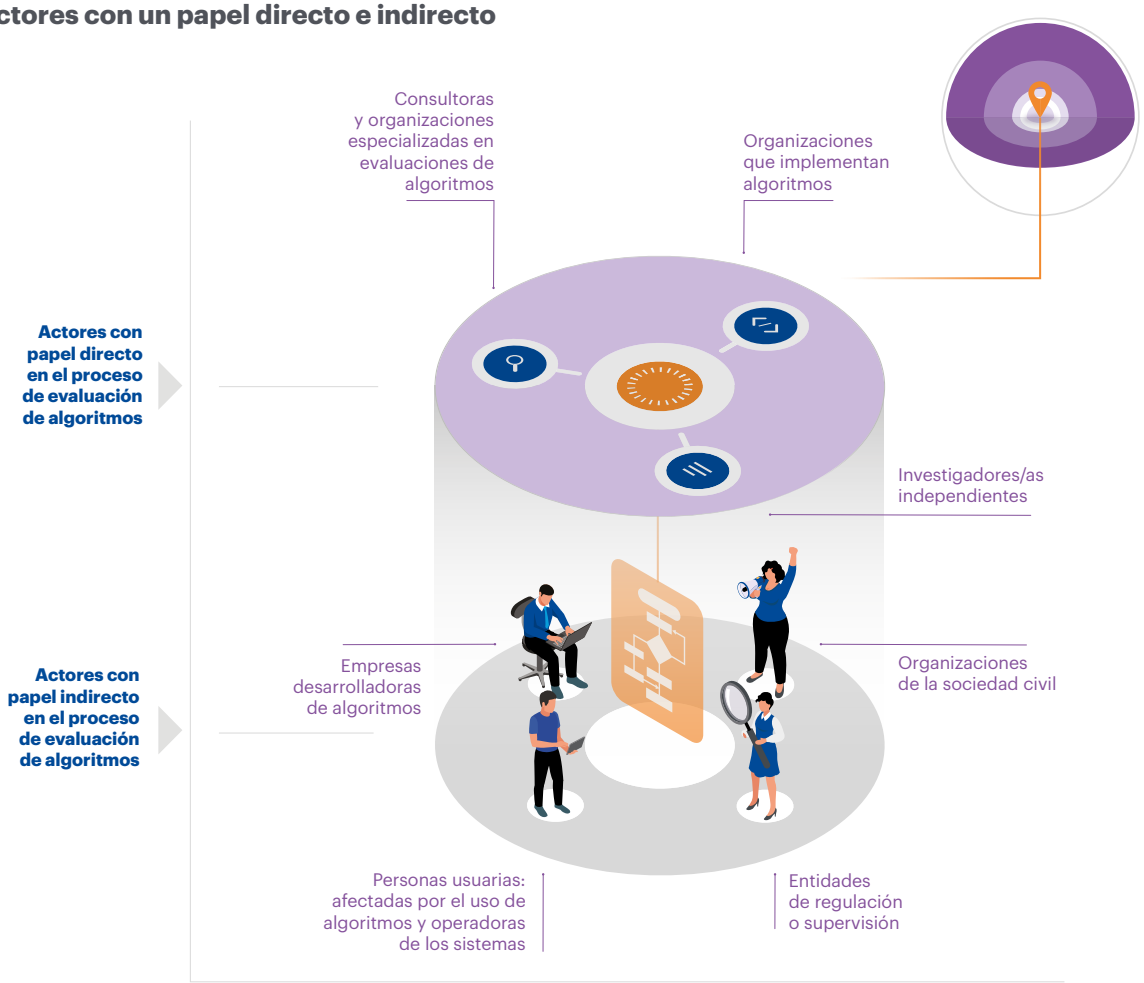
[14] Centre for Data Ethics and Innovation y sus Techniques for assuring AI systems: https://cdeiuk.github.io/ai-assurance-guide/techniques/.

[15] Bajo el liderazgo de Infocomm Media Development Authority, un caso de interés puede ser la AI Verify Foundation (https://aiverifyfoundation.sg), que incluye empresas que implementan IA, desarrolladores de soluciones, usuarios o tomadores de decisiones en el sector público.

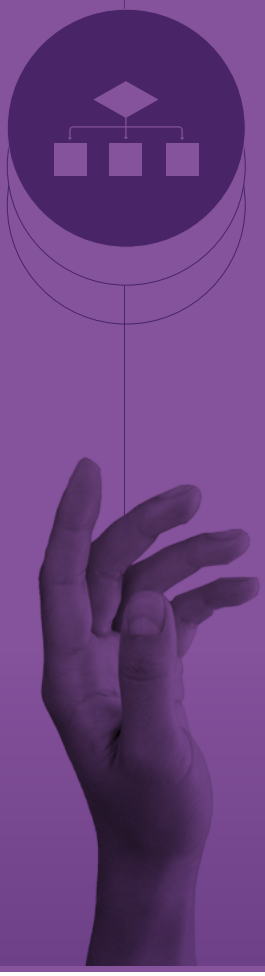
Finalmente, otras **organizaciones de la sociedad civil** o personas defensoras de los derechos humanos y los derechos digitales, grupos de defensa de colectivos con discapacidades, medios de comunicación y periodistas, juristas, profesorado universitario, etc., también pueden contar con un papel al final de este tipo de procesos. Más que involucrarse directamente, parece que estos actores están llamados a formar parte del ecosistema de evaluación de algoritmos colaborando, por ejemplo, en la difusión de información pública, la divulgación social de los impactos de los algoritmos de empresas privadas o el análisis experto en medios de comunicación, entornos académicos, etc.

En conclusión, las dinámicas de rendición de cuentas algorítmica gozarán de mayor solidez técnica, respaldo institucional y confianza social en la medida que se pueda transparentar y poner a disposición del público la máxima cantidad de datos y evidencias posible, mejorando también con ello la sensibilización social ante los nuevos desafíos y oportunidades que este tipo de tecnologías entrañan en el momento actual.

Figura 4.
Nivel *micro* de gobernanza.
Actores con un papel directo e indirecto



Fuente de los datos: elaboración



5. Una mirada hacia el futuro: mejora de los procesos de evaluación de algoritmos

Este documento se completa con una sección orientada a mencionar una serie de puntos prácticos de cara a mejorar los procesos de evaluación de algoritmos. A partir de la información aportada en la revisión bibliográfica y documental y las entrevistas realizadas a personas expertas, así como desde la propuesta de tipología de dimensiones en evaluaciones de algoritmos, los métodos aplicables y los diferentes niveles del ecosistema de rendición de cuentas algorítmica, a continuación, se concretan seis propuestas de mejora para el futuro de este tipo de procesos:

1. **Explorar nuevas estrategias metodológicas para el desarrollo de evaluaciones de algoritmos.**
   En primer lugar, es recomendable aplicar con más frecuencia métodos cualitativos para obtener detalles sobre las experiencias y percepciones en torno a los algoritmos, tanto de las personas que trabajan en las organizaciones como del público general. Sería ideal, además, combinar estos métodos con aquellos más técnicos que se emplean generalmente en este tipo de evaluaciones. De esta forma, se podrán entender los sistemas algorítmicos en toda su complejidad, y ofrecer así medidas de prevención y mitigación más holísticas.

2. **Conformar equipos de evaluación algorítmica que sean multidisciplinares y diversos.**
   La variedad de enfoques y métodos que existen debe motivar a la creación de equipos multidisciplinares, en los que se combinen los conocimientos técnicos y la formación en ciencias sociales. Dependiendo del foco de la evaluación y del sector, se podrían necesitar especialistas en áreas concretas (por ejemplo, en derechos humanos, salud, transporte, defensa, etc.) que enriquezcan el proceso con su visión experta. Por otro lado, se debe fomentar también la diversidad sociodemográfica y cultural del equipo, para incorporar distintas miradas en función de las experiencias personales y colectivas.

3. **Fortalecer el rol de las personas usuarias y las organizaciones de la sociedad civil en el conocimiento sobre el impacto de los algoritmos.**
   En línea con el punto anterior, existe un potencial de implicar a la sociedad en general y a los colectivos más afectados por las decisiones algorítmicas. Algunas herramientas explicadas en secciones anteriores (como IndieLabel, por ejemplo) ofrecen ideas sobre cómo involucrar a usuarios/as reales y comunidades potencialmente perjudicadas en los procesos de evaluación algorítmica, tanto para la reflexión crítica sobre los constructos y supuestos teóricos que están detrás de los algoritmos como para entender el impacto de estos sistemas en determinados colectivos. Es indispensable, en este sentido, que todo el proceso de evaluación esté orientado hacia las personas.

4. **Promover la responsabilidad social corporativa en el sector empresarial ligado a la IA.**
   Es importante que las empresas prioricen el uso responsable de la IA, específicamente con el diseño e implementación de algoritmos éticos y algoritmos verdes, es decir, que tengan en cuenta aspectos relacionados con la ética y los derechos humanos, así como la sostenibilidad ambiental desde el principio del ciclo de vida de la IA[16]. La Agenda 2030 y sus 17 Objetivos de Desarrollo Sostenible (ODS) ofrecen una buena hoja de ruta para alcanzar este fin.

   La idea, en este sentido, es que se genere una sinergia entre varios ámbitos de responsabilidad social: por un lado, que los sistemas algorítmicos que se utilicen estén alineados con esta visión y, por otro lado, que la propia IA pueda servir de vehículo para alcanzar objetivos en beneficio de la sociedad. En este caso, las evaluaciones son fundamentales para garantizar que, en efecto, los algoritmos impulsados desde las empresas privadas sigan estos estándares, siempre con una perspectiva de colaboración y acompañamiento.

5. **Fortalecer la labor de regulación y supervisión del sector público, así como de implementación responsable de algoritmos.**
   En el variado ecosistema de las evaluaciones algorítmicas, el sector público debe ocupar un lugar muy relevante. Es cierto que, como se ha señalado previamente, cada contexto tiene sus dinámicas particulares, pero muchas personas expertas coinciden en que el sector público debe tener un rol fundamental para la definición de estándares y la regulación de estos procesos.

   Con el liderazgo del sector público, se deben impulsar el debate y la colaboración entre distintos sectores para generar normativas y documentos con principios básicos y crear organismos de supervisión, entre otras medidas que favorezcan el equilibrio entre la innovación tecnológica y el control, así como las buenas prácticas profesionales en el ámbito concreto de las evaluaciones algorítmicas. Será necesario, igualmente, debatir sobre las concesiones que se deben hacer y la armonización o los cambios que se deben introducir en la normativa preexistente, en caso de que sea necesario.

[16] El Plan Nacional de Algoritmos Verdes de España apunta en esa dirección: https://portal.mineco.gob.es/RecursosNoticia/mineco/prensa/noticias/2022/20221213_plan_algoritmos_verdes.pdf

6. **Impulsar la maduración de un ecosistema de evaluación de algoritmos, integrando niveles y una perspectiva internacional.**
   Como se ha señalado en secciones anteriores, el incipiente ecosistema de las evaluaciones algorítmicas ha ido creciendo de forma veloz, pero no necesariamente de manera coordinada. Es importante que se fortalezcan las relaciones entre los distintos actores y sectores involucrados, de manera que se pueda avanzar en la definición de unos estándares y principios apropiados para estos procesos. La implicación de empresas, gobiernos, organizaciones de la sociedad civil, universidades, medios de comunicación y público general permitirá impulsar y enriquecer el debate sobre temas urgentes, como la transparencia de las propias evaluaciones, los potenciales conflictos de intereses, la integridad en la práctica profesional de la evaluación algorítmica, etc.

   En resumen, contribuir a que florezca un ecosistema de evaluación de algoritmos de carácter internacional puede ser decisivo para consolidar visiones compartidas, que superen las fronteras nacionales o regionales. Pero también se ha de ir más allá de los límites de la gobernanza de este tipo de procesos considerando los sectores (público, privado y social), las áreas de actividad y los actores participantes, que deberán integrarse lo más armoniosamente posible, en cada caso, para avanzar conjuntamente y mejorar los resultados de estos procesos en beneficio de las personas.

Conclusiones

Este trabajo responde a la necesidad de conocer las **implicaciones de las evaluaciones de algoritmos en un contexto marcado por su aplicación creciente en diferentes ámbitos de la vida humana**. En este documento se han presentado conceptos clave, herramientas y métodos para desarrollar evaluaciones de algoritmos con distintas perspectivas. Se ha abordado, igualmente, el ecosistema de actores y sectores involucrados en estos procesos, con el objetivo de entender este tema en toda su complejidad, y se han presentado seis propuestas de mejora para avanzar hacia evaluaciones algorítmicas que aporten valor a las sociedades actuales.

Particularmente, la pregunta que ha guiado este trabajo, que cuenta con una clara orientación práctica, es la siguiente: **¿cómo se pueden evaluar los algoritmos para detectar los potenciales problemas que contienen y/o que se derivan de su uso, así como contribuir a su mitigación?** A lo largo del documento, se ha destacado la importancia de ampliar la mirada para entender los procesos de evaluación de algoritmos desde una perspectiva holística, con el fin de mitigar riesgos y, por otra parte, aumentar sus beneficios para el bien común.
En este sentido, es necesario combinar el conocimiento puramente tecnológico con enfoques provenientes de disciplinas como el derecho, la sociología, la ciencia política, la psicología, la filosofía, etc., para así abordar el impacto de los sistemas algorítmicos teniendo en cuenta todas las aristas posibles. Solo de esa manera se podrá avanzar hacia el diseño e implementación de algoritmos de una manera responsable, que respete los principios éticos y los derechos de las personas y organizaciones.

Para alcanzar este objetivo, se debe **apostar por un ecosistema de rendición de cuentas algorítmico que aúne los esfuerzos del sector público, el sector privado y las organizaciones del tercer sector**, así como por un trabajo más cooperativo entre distintos ámbitos de actuación en los que los algoritmos pueden tener un impacto relevante. Las personas usuarias, tanto dentro como fuera de las organizaciones, también deben tener un rol importante en los procesos de evaluación, y muy especialmente aquellos colectivos afectados por las decisiones algorítmicas, así como quienes sufren una mayor vulnerabilidad. Este enfoque más abierto, participativo y colaborativo de las evaluaciones debería conducir a cambios fundamentales en el propio diseño e implementación de algoritmos, con la pretensión de que en todo su ciclo de vida las personas ocupen un lugar central.

Referencias

Ada Lovelace Institute. (2020). Examining the Black Box. Tools for assessing algorithmic systems. [PDF] Disponible en: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/ (Consultado: 13-11-2023)

Baeza-Yates, R. y Matthews, J. (2022). Declaración de principios para sistemas algorítmicos responsables. ACM. [PDF] Disponible en: https://www.acm.org/binaries/content/assets/public-policy/spanish-statement-ai.pdf (Consultado: 13-11-2023)

Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. Proceedings of the ACM on Human-Computer Interaction 5(CSCW1), 74:1-74:34. [PDF] Disponible en: https://dl.acm.org/doi/10.1145/3449148 (Consultado: 13-11-2023)

Basu, T., Brennan, J., Kak, A. y Joshi, D. (2021). Algorithmic accountability for the public sector. Learning from the first wave of policy implementation. Ada Lovelace Institute, AI Now y Open Government Partnership. [PDF] Disponible en: https://www.opengovpartnership.org/wp-content/uploads/2021/08/executive-summary-algorithmic-accountability.pdf (Consultado: 13-11-2023)

Baykurt, B. (2022). Algorithmic accountability in U.S. cities: Transparency, impact, and political economy. Big Data & Society 9(2). [PDF] Disponible en: https://journals.sagepub.com/doi/10.1177/20539517221115426 (Consultado: 13-11-2023)

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. European Law Journal 13(4), 447-468. [PDF] Disponible en: https://onlinelibrary.wiley.com/doi/epdf/10.1111/j.1468-0386.2007.00378.x (Consultado: 13-11-2023)

Brown, S., Davidovic, J. y Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. Big Data & Society 8(1). [PDF] Disponible en: https://journals.sagepub.com/doi/10.1177/2053951720983865 (Consultado: 13-11-2023)

Buolamwini, J. y Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research 81. [PDF] Disponible en: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf (Consultado: 13-11-2023)

Burr, C. y Leslie, D. (2022). Ethical Assurance: A Practical Approach to the Responsible Design, Development, and Deployment of Data-Driven Technologies. AI and Ethics 3(1), 1-26. [online] Disponible en: https://link.springer.com/article/10.1007/s43681-022-00178-0 (Consultado: 13-11-2023)

Costanza-Chock, S., Raji, I. D. y Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 1571-1583. [PDF] Disponible en: https://dl.acm.org/doi/10.1145/3531146.3533213 (Consultado: 13-11-2023)

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. [online] Disponible en: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (Consultado: 13-11-2023)

De Manuel, A., Delgado, J., Parra Jounou, I., Ausín, T., Casacuberta, D., Cruz, M., Guersenzvaig, A., Moyano, C., Rodríguez-Arias, D., Rueda, J. y Puyol, A. (2023). Ethical assessments and mitigation strategies for biases in AI-systems used during the COVID-19 pandemic. Big Data & Society 10(1). [online] Disponible en: https://journals.sagepub.com/doi/10.1177/20539517231179199 (Consultado: 13-11-2023)

DeVito, M. A. (2017). From Editors to Algorithms. Digital Journalism 5(6), 753-773. [PDF] Disponible en: https://www.tandfonline.com/doi/full/10.1080/21670811.2016.1178592 (Consultado: 13-11-2023)

DeVos, A., Dhabalia, A., Shen, H., Holstein, K. y Eslami, M. (2022). Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. CHI Conference on Human Factors in Computing Systems 1-19. [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3491102.3517441 (Consultado: 13-11-2023)

Dodge, J., Khanna, R., Irvine, J., Lam, K., Mai, T., Lin, Z., Kiddle, N., Newman, E., Anderson, A., Raja, S., Matthews, C., Perdriau, C., Burnett, M. y Fern, A. (2021). After-Action Review for AI (AAR/AI). ACM Transactions on Interactive Intelligent Systems 11(3-4), 29:1-29:35. [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3453173 (Consultado: 13-11-2023)

Eriksson, M. y Johansson, A. (2017). Tracking Gendered Streams. Culture Unbound, 9(2), 163-183. [PDF] Disponible en: https://www.diva-portal.org/smash/get/diva2:1243114/FULLTEXT01.pdf (Consultado: 13-11-2023)

Eticas Consulting. (s.f.). ¿Cómo se audita un algoritmo? Los cinco pasos de una auditoría algorítmica. [online]. Disponible en: https://www.eticasconsulting.com/como-se-audita-un-algoritmo-pasos-para-auditar-algoritmos/ (Consultado: 13-11-2023)

Eubanks, V. (2019). La automatización de la desigualdad. Herramientas de tecnología avanzada para supervisar y castigar a los pobres. Madrid, España: Capitán Swing.

Galdon Clavell, G., Martín Zamorano, M., Castillo, C., Smith, O. y Matic, A. (2020). Auditing Algorithms: On Lessons Learned and the Risks of Data Minimization. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 265-271. [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3375627.3375852 (Consultado: 13-11-2023)

Garde Roca, J. A. (2023). ¿Pueden los algoritmos ser evaluados con rigor? Encuentros Multidisciplinares, 73, 1-13. [PDF] Disponible en: http://www.encuentros-multidisciplinares.org/revista-73/juan-antonio-garde.pdf (Consultado: 13-11-2023)

Godin, K., Stapleton, J., Kirkpatrick, S. I., Hanning, R. M. y Leatherdale, S. T. (2015). Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada. Systematic Reviews 4(1), 138. [PDF] Disponible en: https://systematicreviewsjournal.biomedcentral.com/articles/10.1186/s13643-015-0125-0 (Consultado: 13-11-2023)

Goodman, E. P. y Trehu, J. (2022). AI Audit-Washing and Accountability. GMF. [PDF] Disponible en: https://www.gmfus.org/news/ai-audit-washing-and-accountability (Consultado: 13-11-2023)

Groves, L. (2022). Algorithmic impact assessment: a case study in healthcare. Ada Lovelace Institute. [PDF] Disponible en: https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/ (Consultado: 13-11-2023)

Hamilton, M. (2021). Evaluating Algorithmic Risk Assessment. New Criminal Law Review 24(2), 156-211. [PDF] Disponible en: https://online.ucpress.edu/nclr/article/24/2/156/116809/Evaluating-Algorithmic-Risk-Assessment (Consultado: 13-11-2023)

Kelly-Lyth, A. y Thomas, A. (2023). Algorithmic management: Assessing the impacts of AI at work. European Labour Law Journal 14(2), 230-252. [PDF] Disponible en: https://journals.sagepub.com/doi/10.1177/20319525231167478 (Consultado: 13-11-2023)

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, Gregorovic, M., Khan, S. y Lomas, E. (2021). Towards Algorithm Auditing. A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms. SSRN Electronic Journal. [PDF] Disponible en: https://discovery.ucl.ac.uk/id/eprint/10164738/1/owards%20Algorithm%20Auditing%20A%20Survey%20on%20Managing%20Legal,%20Ethical%20and%20Technological%20Risks%20of%20AI,%20ML%20and%20Associated%20Algorithms.pdf (Consultado: 13-11-2023)

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P. y Karahalios, K. (2019). Search bias quantification: investigating political bias in social media and web search. Information Retrieval Journal 22(1), 188-227. [PDF] Disponible en: https://link.springer.com/article/10.1007/s10791-018-9341-2 (Consultado: 13-11-2023)

Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P. y Metaxa, D. (2023). Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. Proceedings of the ACM Human-Computer Interaction [PDF] Disponible en: https://hci.stanford.edu/publications/2023/Lam_STA_CSCW23.pdf (Consultado: 13-11-2023)

Mantelero, A. (2018). AI and Big Data: A blueprint for a human rights, social and ethical impact assessment. Computer Law & Security Review 34(4), 754-772. [online] Disponible en: https://www.sciencedirect.com/science/article/pii/S0267364918302012?via%3Dihub (Consultado: 13-11-2023)

Meßmer, A.-K. y Degeling, M. (2023). Auditing Recommender Systems. Putting the DSA into practice with a risk-scenario-based approach. [PDF] Stiftung Neue Verantwortung. Disponible en: https://www.stiftung-nv.de/de/publication/auditing-recommender-systems (Consultado: 13-11-2023)

Metcalf, J., Moss, E., Watkings, E. A., Singh, R. y Elish, M. C. (2021). Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3442188.3445935 (Consultado: 13-11-2023)

Minkkinen, M., Laine, J. y Mäntymäki, M. (2022). Continuous Auditing of Artificial Intelligence: A Conceptualization and Assessment of Tools and Frameworks. Digital Society 1(3), 21. [PDF] Disponible en: https://link.springer.com/article/10.1007/s44206-022-00022-2 (Consultado: 13-11-2023)

Mökander, J., Axente, M., Casolari, F. y Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. Minds and Machines 32(2), 241-268. [PDF] Disponible en: https://link.springer.com/article/10.1007/s11023-021-09577-4 (Consultado: 13-11-2023)

Morondo, D. y Eguiluz, J. A. (2022). La discriminación algorítmica en España: límites y potencial del marco legal. Digital Future Society Think Tank. [PDF] Disponible en: https://digitalfuturesociety.com/es/report/algorithmic-discrimination-in-spain/ (Consultado: 13-11-2023)

Munz, P., Hennick, M. y Stewart, J. (2023). Maximizing AI reliability through anticipatory thinking and model risk audits. AI Magazine 44(2), 173-184. [PDF] Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/aaai.12099 (Consultado: 13-11-2023)

Novelli, C., Casolari, F., Rotolo, A., Taddeo, M. y Floridi, L. (2023). Taking AI risks seriously: A new assessment model for the AI Act. AI & SOCIETY. [PDF] Disponible en: https://link.springer.com/article/10.1007/s00146-023-01723-z (Consultado: 13-11-2023)

Oswald, M., Grace, J., Urwin, S. y Barnes, G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. Information & Communications Technology Law 27(2), 223-250. [PDF] Disponible en: https://www.tandfonline.com/doi/epdf/10.1080/13600834.2018.1458455?needAccess=true (Consultado: 13-11-2023)

Papakyriakopoulos, O. y Mboya, A. M. (2023). Beyond Algorithmic Bias: A Socio-Computational Interrogation of the Google Search by Image Algorithm. Social Science Computer Review 41(4), 1100-1125. [PDF] Disponible en: https://journals.sagepub.com/doi/10.1177/08944393211073169 (Consultado: 13-11-2023)

Pappu, A., Brennan, J., Strait, A., Parker, I. y Jones, E. (2021). Technical methods for regulatory inspection of algorithmic systems in social media platforms (Ethics and accountability in practice). Ada Lovelace Institute. [PDF] Disponible en: www.adalovelaceinstitute.org/wp-content/uploads/2021/12/ADA_Technical-methods-regulatory-inspection_report.pdf (Consultado: 13-11-2023)

Percy, C., Dragicevic, S., Sarkar, S. y d'Avila Garcez, A. (2021). Accountability in AI: From principles to industry-specific accreditation. AI Communications 34(3), 181-196. [PDF] Disponible en: https://content.iospress.com/articles/ai-communications/aic210080 (Consultado: 13-11-2023)

Raji, I. D. y Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society 429-435. [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3306618.3314244 (Consultado: 13-11-2023)

Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T. y Ghani, R. (2019). Aequitas: A Bias and Fairness Audit Toolkit. arXiv. [PDF] Disponible en: https://arxiv.org/abs/1811.05577 (Consultado: 13-11-2023)

Sandu, I., Wiersma, M. y Manichand, D. (2022). Time to audit your AI algorithms. Maandblad Voor Accountancy En Bedrijfseconomie 96(7/8). [PDF] Disponible en: https://mab-online.nl/article/90108/ (Consultado: 13-11-2023)

Sandvig, C., Hamilton, K., Karahalios, K. y Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. University of Michigan. [PDF] Disponible en: https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf (Consultado: 13-11-2023)

Schellmann, H. (2021). Auditors are testing hiring algorithms for bias, but there's no easy fix. MIT Technology Review. [online] Disponible en: https://www.technologyreview.com/2021/02/11/1017955/auditors-testing-ai-hiring-algorithms-bias-big-questions-remain/ (Consultado: 13-11-2023)

Sheehan, M. y Du, S. (2022). What China's Algorithm Registry Reveals about AI Governance. Carnegie Endowment for International Peace. [online] Disponible en: https://carnegieendowment.org/2022/12/09/what-china-s-algorithm-registry-reveals-about-ai-governance-pub-88606 (Consultado: 13-11-2023)

Sloane, M. (2021). The Algorithmic Auditing Trap. OneZero. [online] Disponible en: https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d (Consultado: 13-11-2023)

Spyridou, P. (Lia), Djouvas, C. y Milioni, D. (2022). Modeling and Validating a News Recommender Algorithm in a Mainstream Medium-Sized News Organization: An Experimental Approach. Future Internet 14(10) [PDF] Disponible en: https://www.mdpi.com/1999-5903/14/10/284 (Consultado: 13-11-2023)

Srba, I., Moro, R., Tomlein, M., Pecher, B., Simko, J., Stefancova, E., Kompan, M., Hrckova, A., Podrouzek, J., Gavornik, A. y Bielikova, M. (2023). Auditing YouTube's Recommendation Algorithm for Misinformation Filter Bubbles. ACM Transactions on Recommender Systems 1(1). [online] Disponible en: https://dl.acm.org/doi/10.1145/3568392 (Consultado: 13-11-2023)

Stahl, B. C., Antoniou, J., Bhalla, N., Brooks, L., Jansen, P., Lindqvist, B., Kirichenko, A., Marchal, S., Rodrigues, R., Santiago, N., Warso, Z. y Wright, D. (2023). A systematic review of artificial intelligence impact assessments. Artificial Intelligence Review 56(11). [PDF] Disponible en: https://link.springer.com/article/10.1007/s10462-023-10420-8 (Consultado: 13-11-2023)

Stanford University Human-Centered Artificial Intelligence. (2023). Artificial Intelligence Index Report 2023. [PDF] Disponible en: https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf (Consultado: 13-11-2023)

Strubell, E., Ganesh, A. y McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. arXiv. [PDF] Disponible en: https://arxiv.org/abs/1906.02243 (Consultado: 13-11-2023)

Wasilow, S. y Thorpe, J. B. (2019). Artificial Intelligence, Robotics, Ethics, and the Military: A Canadian Perspective. AI Magazine 40(1), 37-48. [PDF] Disponible en: https://onlinelibrary.wiley.com/doi/10.1609/aimag.v40i1.2848 (Consultado: 13-11-2023)

Wilson, C., Mislove, A., Ghosh, A. y Jiang, S. (2020). Auditing the pymetrics Model Generation Process. [PDF] Disponible en: https://cbw.sh/static/audit/pymetrics/pymetrics_audit_result_whitepaper.pdf (Consultado: 13-11-2023)

Yam, J. y Skorburg, J. A. (2021). From human resources to human rights: Impact assessments for hiring algorithms. Ethics and Information Technology 23(4), 611-623. [online] Disponible en: https://link.springer.com/article/10.1007/s10676-021-09599-7 (Consultado: 13-11-2023)

Anexo

Metodología

Para la elaboración de este informe, se ha desarrollado un trabajo de campo compuesto por tres fases principales: a) revisión sistemática de literatura académica, b) análisis documental de literatura gris, y c) entrevistas en profundidad a especialistas y actores clave en los procesos de evaluación algorítmica, tanto en España, Reino Unido y Holanda como en Estados Unidos. A continuación, explicamos los detalles más relevantes de cada fase.

La pregunta de investigación que ha guiado tanto la revisión de literatura académica como la de la literatura gris es la siguiente: ¿cuáles son las principales herramientas y metodologías de evaluación de algoritmos que se pueden encontrar en la literatura? El objetivo principal de estas fases del trabajo de campo fue identificar y sistematizar la información que se ha publicado hasta el momento sobre el tema.

Revisión sistemática de literatura académica
En esta fase, se revisaron artículos científicos publicados en revistas indexadas en las bases de datos JCR y SJR, así como ponencias en conferencias relevantes. La búsqueda se realizó en la base de datos Web of Science, en julio del 2023, con la siguiente secuencia de términos: TI = ("algorithm*" OR "AI" OR "artificial intelligence" OR "automated system*" OR "machine learning" OR "deep learning") AND TS = ("audit*" OR "assessment*"). Es decir, se tomaron en cuenta aquellos artículos que incluían en el título términos relacionados con la IA, y cuyo tema se relacionara con las evaluaciones.

En cuanto a la fecha de publicación, se consideraron todos los estudios incluidos en la base de datos desde 1985 hasta el 2023, en las siguientes áreas de investigación: ciencias de la computación (concretamente, sistemas de información e IA), ciencias de la información, gestión, economía, ciencias sociales, negocios, derecho, sociología y Administración pública. La búsqueda con estos parámetros arrojó un total de 2.907 documentos.

Una vez obtenidos estos resultados, se descargó la base de datos con la información relevante de cada artículo y se definieron los criterios de inclusión. En este caso, se consideraron para el análisis final únicamente aquellos artículos que abordaban de forma central el tema de las evaluaciones de algoritmos, independientemente del sector, o que incluían en su metodología alguna información relevante sobre métodos para hacer evaluaciones algorítmicas.

Posteriormente, se desarrolló una etapa de revisión de títulos y resúmenes, con el fin de seleccionar los textos más pertinentes para la síntesis final del contenido. Para desarrollar este proceso, se usó ASReview (https://asreview.nl/), una herramienta de *active learning* y de código abierto que permite agilizar los procesos de revisión sistemática de literatura. Al utilizarla, la persona investigadora va etiquetando los artículos como relevantes o irrelevantes, y el modelo se va entrenando progresivamente para identificar de forma prioritaria los documentos de mayor interés. De este modo, no es necesario revisar manualmente la base de datos completa para encontrar los textos que se necesitan.

Durante dicha etapa, se encontraron 62 artículos relevantes. Para compensar las posibles fallas del método anterior, se complementó el proceso con una búsqueda manual de artículos en la Web y con las recomendaciones de las personas entrevistadas. De esta forma, la base de datos aumentó a 90 documentos. En el proceso de la revisión más detallada de los artículos, quedaron excluidos 26 por no cumplir con los criterios específicos de búsqueda y análisis de la información. En este sentido, el total de artículos incluidos en la síntesis final fue de 64.

Revisión sistemática de informes y otros tipos de publicaciones
Para completar la información obtenida en la revisión de literatura académica, se desarrolló también una fase de búsqueda y análisis de documentos de literatura gris: informes y otros tipos de publicaciones (posts de blogs, páginas web, etc.) tanto de organismos públicos como de organizaciones del tercer sector, universidades, *think tanks*, empresas y otras entidades. En este caso, se hizo una búsqueda convencional en Google, siguiendo algunas pautas del método explicado por Godin et al. (2015). Se utilizaron varias secuencias de términos, que se detallan a continuación:

- audit + algorithm
- audit + artificial intelligence
- assessment + algorithm
- assessment + artificial intelligence
- auditoría + algoritmo
- auditoría algorítmica
- evaluación de algoritmos
- evaluación algorítmica

Se revisaron los primeros 100 resultados de cada una de las secuencias de términos. Específicamente, se hizo una lectura rápida de cada título y resumen para verificar si los documentos abordaban de forma específica el tema de las auditorías y evaluaciones de algoritmos. En caso de que fuese así, se incluyeron directamente en la base de datos para la posterior extracción de datos y síntesis de resultados.

Para evitar en lo posible la pérdida de resultados relevantes y siguiendo estudios previos (Godin et al. 2015), esta búsqueda se complementó con la revisión manual de las páginas web de algunos organismos de referencia sobre IA y algoritmos (como Ada Lovelace Institute, AI Now Institute, AI Watch, Stanford University Human-Centered Artificial Intelligence, OCDE y otros organismos europeos, etc.). En la Tabla 4 se detalla la cantidad de documentos incluidos en la revisión.

Tabla 4.
Documentos de literatura gris de cada tipo revisados

Tipo de documento	Número de documentos
Documentos oficiales	11
Informes	18
Posts de blogs y publicaciones en páginas web	23
Libros y capítulos de libros	3
Textos normativos	5
Total	**60**

Fuente de los datos: elaboración propia

Realización de entrevistas a personas expertas

La tercera fase del trabajo de campo se desarrolló entre agosto y septiembre del 2023, y consistió en llevar a cabo quince entrevistas semiestructuradas a personas especialistas en IA y evaluaciones algorítmicas, que trabajan en organismos internacionales, organismos europeos, empresas y consultoras privadas, universidades y organizaciones del tercer sector (véase la Tabla 5). Todas las entrevistas se desarrollaron en formato online, a través de Google Meet o Zoom, nueve de ellas en inglés y seis en español.

Tabla 5.
Personas entrevistadas según el organismo u organización de procedencia

Tipo de organismo	Número de personas
Organismo independiente de investigación	1
Organización del tercer sector	3
Empresas y consultoras privadas	2
Organismo europeo o internacional	4
Institución académica	3
Investigadora independiente	2
Total	**15**

Fuente de los datos: elaboración propia

El objetivo de las entrevistas era obtener información conceptual y práctica sobre el desarrollo de evaluaciones algorítmicas en diferentes contextos, para validar y contrastar los datos obtenidos durante la revisión sistemática de literatura. En este sentido, se siguió un protocolo compuesto por 14 preguntas, que podían variar en función de la dinámica de la conversación.

Cuestionario de partida

1. Por favor, díganos su nombre y puesto actual en su organismo u organización.
2. ¿Cuáles son sus responsabilidades actuales? ¿Cómo se relacionan con la inteligencia artificial en general y las evaluaciones de algoritmos en particular?
3. ¿Cuál cree que es la mejor definición de auditoría algorítmica?
4. ¿Cuáles son los principales tipos de auditorías y evaluaciones algorítmicas? ¿Y cuáles son sus diferencias y similitudes?
5. ¿Cuáles son las metodologías y herramientas de auditoría y evaluación de algoritmos con las que tiene experiencia? ¿Cuáles son sus principales características? ¿Podría mencionar un ejemplo específico?
6. ¿Podría describir el proceso seguido por su organismo u organización para realizar auditorías y evaluaciones de algoritmos? (es decir, quién participa, cómo define las estrategias, cómo comunica los resultados, etc.)
7. ¿Cuál es el marco legal e institucional definido en su contexto para realizar auditorías y evaluaciones de algoritmos?
8. ¿Cómo cree que el sector público debería realizar auditorías y evaluaciones de algoritmos?
9. ¿Conoce casos de Administraciones públicas que actualmente realizan este tipo de auditorías?
10. Según su experiencia, ¿cuáles son las principales lecciones aprendidas con respecto a las auditorías y evaluaciones de algoritmos? (es decir, desafíos, oportunidades...)
11. ¿Qué considera que se debe hacer para mejorar estos procesos en el futuro?
12. ¿Le gustaría agregar algún comentario?
13. ¿Tiene acceso a algún informe o documento que pueda brindar información relevante para nuestra investigación?
14. En su opinión, ¿deberíamos considerar a otra persona clave que cree que podría ser entrevistada sobre este mismo tema?

Agradecimientos

Autores

J. Ignacio Criado es profesor titular en el Departamento de Ciencia Política y Relaciones Internacionales, y Director del Lab Innovación, Tecnología y Gestión Pública, en la Universidad Autónoma de Madrid. Sus intereses se centran en temas de gobierno abierto, administración digital, innovación pública y redes sociales, así como gobernanza algorítmica e inteligencia artificial en el sector público.

Ariana Guevara-Gómez es profesora ayudante en el Departamento de Ciencia Política y Relaciones Internacionales, e investigadora en el Lab Innovación, Tecnología y Gestión Pública, de la Universidad Autónoma de Madrid. Investiga sobre género, tecnología e inteligencia artificial, innovación y administración pública.

Coordinación de investigación

J. Ignacio Criado

Coordinación del proyecto

Tanya Álvarez dirige la investigación de Digital Future Society Think Tank sobre brechas digitales y digitalización del sector público. Aboga por una perspectiva interdisciplinar del impacto de la tecnología en la sociedad. Es graduada en Historia del Arte por el Swarthmore College y tiene un máster en Gestión del Patrimonio Cultural por la Universidad de Barcelona.

Edición y diseño

Marta Campo, editora y correctora

Manuela Moulian, diseñadora y autora de las infografías

Personas entrevistadas

Adriano Soares Koshiyama, cofundador de la empresa Holistic AI

Albert Sabater, profesor de Sociología Computacional en la Universitat de Girona y director del Observatorio de Ética en Inteligencia Artificial de Cataluña

Aparna Surendra, directora del equipo de Investigación y Conocimiento de AWO Agency (con sedes en Londres, Bruselas y París)

Carlos Castillo, profesor ICREA en la Universidad Pompeu Fabra de Barcelona y coordinador del Grupo Ciencias de la Web e Informática Social

Dafna Feinholz, directora de la sección de Bioética y Ética de la Ciencia de la UNESCO

Fabio Curi, desarrollador *back-end* y consultor de IA en la OCDE

Ibán García del Blanco, diputado español en el Parlamento Europeo

Javier de la Cueva, patrono de la Fundación Civio y especialista en Derecho y Tecnologías de la Información y la Comunicación

Jurriaan Parie, integrante del Consejo de Algorithm Audit

Krishnaram Kenthapadi, Chief AI Officer y Chief Scientist de Fiddler AI

Lara Groves, investigadora en el Ada Lovelace Institute, en el Reino Unido

Ola Al Khatib, investigadora predoctoral en la Universidad de Utrecht e integrante del equipo de Algorithm Audit

Rashad Abelson, Technology Sector Lead y Due Diligence Legal Expert en la OCDE

Ricardo Baeza-Yates, director de investigación del Institute for Experiential AI de la Northeastern University, en Estados Unidos

Shazade Jameson, investigadora y consultora independiente de gobernanza de la IA, específicamente en temas urbanos

Este informe se debe citar de la siguiente manera:
Digital Future Society (2024). Hacia un uso responsable de los algoritmos: métodos y herramientas para su auditoría y evaluación

Datos de contacto:
thinktank@digitalfuturesociety.com

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
red.es
Mobile WorldCapital Barcelona