Digital Future Society

Hacia una supervisión significativa de los sistemas automatizados de toma de decisiones

Un programa de

GOBIERNO DE ESPAÑA
VICEPRESIDENCIA PRIMERA DE GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

MOBILE WORLD CAPITAL 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
VICEPRESIDENCIA PRIMERA DE GOBIERNO
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

red.es

MOBILE WORLD CAPITAL BARCELONA

Publicado
Octubre del 2022

Digital Future Society

Contenidos

1. Introducción

La inteligencia artificial (IA) ha llegado a casi todos los sectores en todo el mundo. Ante su promesa de mejorar la eficiencia y llevar a cabo las tareas repetitivas por medios digitales, el número de Administraciones públicas que emplean sistemas automatizados de toma de decisiones (ADMS) aumenta cada año (Misuraca y Noordt 2020; Zuiderwijk et al. 2021).

Pero esto entraña ciertos riesgos, como el potencial de que los algoritmos preserven desigualdades, sesgos y discriminaciones históricos (Digital Future Society 2020). Además, la automatización de decisiones plantea dudas en torno a la cuestión de la responsabilidad: si un algoritmo toma una decisión que es discriminatoria, ¿quién es el responsable?

Claramente, los Gobiernos se enfrentan a diversos retos al tratar de mejorar la eficiencia automatizando procesos y digitalizando la sociedad, y Digital Future Society (DFS) contribuye desde hace tiempo a los debates relacionados con esos retos.[1] El informe *Gobernanza y algoritmos: riesgos y potencial del uso de la inteligencia artificial en el sector público* analiza los intentos de los Gobiernos de implementar dichos ADMS (Digital Future Society 2021).

Dicho informe se centra en el estudio de los diferentes niveles de algoritmos de gobernanza existentes, y pone de relieve las falsas promesas que acompañan a estos sistemas. En lugar de permitir ser gobernados por la IA, el informe propugna un proceso de gobernanza conjunta con la IA, de modo que las personas ocupen "su posición clásica, es decir, la de utilizar y controlar una tecnología que refuerce su capacidad mediante un proceso que requiera supervisión humana" (Ibid.).

Los ADMS pueden ser problemáticos debido a que pueden reproducir sesgos (preexistentes) que lleven a resultados discriminatorios, causados, por ejemplo, por el diseño de dichos sistemas o por los datos utilizados para entrenarlos. Por ello, es fundamental plantearse cómo podemos asegurarnos de que los ADMS se usen de manera fiable en la Administración pública.

Se espera que la intervención y la supervisión humanas resuelvan este problema controlando las decisiones, teniendo la última palabra o mitigando los errores presentes en los datos o en los resultados del algoritmo. Sin embargo, en la normativa actual, la supervisión se plantea como una medida de seguridad para los sistemas algorítmicos, lo cual es, cuando menos, ambiguo.

1. Véanse los informes de DFS, entre ellos, *Gobernanza y algoritmos: riesgos y potencial del uso de la inteligencia artificial en el sector público* (https://digitalfuturesociety.com/es/report/governing-algorithms/), *Brecha de género en los datos: hacia la igualdad de género en el bienestar digital* (https://digitalfuturesociety.com/es/report/hacia-la-igualdad-de-genero-en-el-estado-de-bienestar-digital/), *Inclusión por diseño: exploración de diseños sensibles al género en el bienestar digital* (https://digitalfuturesociety.com/es/report/exploring-gender-responsive-designs-in-digital-welfare/) y *La confluencia entre la tecnología emergente y el gobierno* (https://digitalfuturesociety.com/es/report/donde-la-tecnologia-emergente-se-encuentra-con-el-gobierno/)

Actualmente, el Reglamento General de Protección de Datos (RGPD) es la única regulación europea que aborda esta cuestión, pero no es un reglamento específico para la inteligencia artificial, sino que se centra principalmente en la protección de datos. El RGPD requiere que los desarrolladores, las empresas, las organizaciones y las Administraciones implementen mecanismos de supervisión humana junto con los algoritmos.

Dichos mecanismos aseguran que una persona actuará en caso de que el algoritmo produzca errores o resultados discriminatorios. Sin embargo, los mecanismos que se proponen son demasiado difusos y no abarcan la complejidad que presentan los casos de uso reales (Green 2021). Además, el reglamento da por supuesto que en los sistemas automatizados no hay ningún tipo de intervención humana cuando, en realidad, pueden existir muchos tipos de interacción humana dentro de esos sistemas, cuyas diferentes implicaciones deben tenerse en cuenta (Binns y Veale 2021).

Dado que el RGPD no es un reglamento específico para la inteligencia artificial, es comprensible que pase por alto las múltiples facetas de esta cuestión. Así pues, sigue existiendo el problema de que la legislación actual no comprende la complejidad inherente a la supervisión humana de los ADMS. No aborda con claridad cuándo ni en qué condiciones puede la supervisión humana ser una respuesta satisfactoria a los sesgos, daños y experiencias problemáticas que se han observado al usar ADMS.

Lamentablemente, lo único que ha quedado claro hasta el momento es que lo que aparenta ser una tarea sencilla (la supervisión humana de las decisiones automatizadas) es mucho más complejo y, a veces, contraproducente (Campolo y Crawford 2020). Un aspecto positivo es que la Unión Europea planea ampliar la regulación de la inteligencia artificial. Pero aún está por ver si la nueva normativa entenderá verdaderamente la complejidad que supone implementar una supervisión humana significativa de los ADMS.

Figura 1. **Supervisión humana de ADMS**

Fuente de la imagen: Digital Future Society.

Acerca de este informe sobre políticas públicas

Este documento trata de analizar la complejidad que hay detrás de la supervisión humana, tanto en su definición como en su regulación y puesta en práctica. Dado que existe un gran debate en torno a cómo debería regularse la interacción entre personas y algoritmos, y en torno si la supervisión humana requerida actualmente basta para mitigar los daños producidos por los algoritmos, este informe sobre políticas públicas contribuye a dicho debate, explorando la normativa y el campo de estudio de la interacción persona-ordenador, a fin de proponer recomendaciones que puedan ayudar a que las personas participen de un modo significativo.

En la primera sección, el documento examina cómo se define la supervisión humana de los sistemas automatizados de toma de decisiones dentro de la regulación de la Unión Europea. Se explica de qué manera la normativa vigente, el RGPD, requiere la supervisión humana y, a continuación, se analiza brevemente la Ley de Inteligencia Artificial, anticipando los aspectos en los que esta propuesta de reglamento podría ofrecer una protección insuficiente.

La segunda sección parte del campo de estudio de la interacción persona-ordenador para explicar mejor los numerosos niveles que abarcan los contextos de las decisiones automatizadas. Se empieza abordando el aspecto humano de la interacción y, en particular, los sesgos que influyen en las personas durante el proceso de toma de decisiones. Luego, se analizan los distintos roles que desempeña la automatización a la hora de ayudar a los operadores humanos en el proceso de toma de decisiones, haciendo referencia a la estructura simplificada propuesta por los investigadores Reuben Binns y Michael Veale.

Empleando dichos tipos simplificados, la tercera sección presenta varios estudios de caso de sistemas automatizados de toma de decisiones en Europa. Mediante el análisis de estos estudios de caso, el documento trata de ofrecer una perspectiva más clara de las circunstancias en que puede ser eficaz o no la supervisión humana.

Finalmente, este informe ofrece una serie de recomendaciones para mejorar la supervisión humana como una herramienta significativa dentro de los ADMS.

¿Por qué ahora?

"Los responsables políticos y las empresas deseosos de encontrar una "solución regulatoria" a los usos dañinos de la tecnología deben reconocer los límites de la supervisión humana y atenerse a ellos, en lugar de presentar la implicación humana —incluso la implicación humana "significativa"— como un antídoto a los daños de los algoritmos. Para ello, es necesario alejarse de las concepciones abstractas de las máquinas y las personas por separado, y pensar en la naturaleza específica de las interacciones persona-algoritmo".[2]
— Ben Green y Amba Kak

Con la creciente adopción de herramientas de automatización en el sector público, los responsables políticos, funcionarios y administradores deben comprender cómo afecta la automatización a los contextos de toma de decisiones. Aunque la normativa suele promover la supervisión humana como una medida de protección, los expertos advierten sobre la falsa sensación de seguridad que promete dicha supervisión, puesto que el riesgo que suponen los ADMS va más allá del criterio del personal de asistencia.

La normativa, hasta ahora, refleja una comprensión superficial de la interacción persona-máquina. Por ello, para minimizar eficazmente los daños por sesgos y discriminación en la toma de decisiones, ya provengan de las personas o de los algoritmos, los responsables políticos deben entender primero los riesgos y la complejidad que comporta el uso de ADMS y de qué manera puede desempeñar un rol significativo la supervisión humana.

Aspectos complejos que deben tenerse en cuenta

Al hablar de supervisión humana, se suele presuponer erróneamente que la intervención humana solo debe producirse en ciertos momentos críticos. En el contexto de los ADMS, el momento crítico suele considerarse el de la decisión final. Este error ha llevado, por ejemplo, a la normativa europea actual, en la que no se contempla el hecho de que el sistema puede tener diferentes relaciones con distintos tipos de usuarios.

Otro factor es que, para tratar de simplificar los procesos, el diseño de muchos sistemas omite por completo la intervención humana: hay sistemas que se usan hoy en día y que no cuentan con mecanismos que garanticen una supervisión eficaz. Otras consideraciones son la posibilidad de que las personas no detecten la influencia que ejercen sobre ellas las máquinas o no perciban las conclusiones incorrectas de estas, o la de que el propio personal que maneja los sistemas sea parcial al enfrentarse a decisiones concretas o carezca de la información, autoridad o comprensión necesarias para intervenir correctamente en los procesos.

Estas son algunas de las tensiones que se han pasado por alto y que la propuesta de reglamento de la IA debería aclarar para que la supervisión humana sea eficaz. Los estudios de caso que se presentan más adelante exploran las consecuencias de no tener en cuenta estos factores.

2. Véase: https://slate.com/technology/2021/06/human-oversight-artificial-intelligence-laws.html

Metodología

El contenido de este informe sobre políticas públicas se basa en una profunda revisión de la literatura de tres campos específicos y en una revisión sistemática de estudios de caso. La revisión de la literatura abordó la comunicación de riesgos, los sistemas de auditoría de algoritmos y la interacción persona-ordenador. En dicha revisión, se prestó especial atención a las publicaciones en revistas de renombre y a las ponencias de congresos sobre la equidad y la ética de los algoritmos.[3]

Para seleccionar los estudios de caso, se exploraron diferentes observatorios de algoritmos[4] para encontrar casos de uso de ADMS en Europa que incluyeran algún tipo de interacción humana. A continuación, los estudios de caso se analizaron según el marco formulado por Reuben Binns y Michael Veale (2021) para elegir casos que ilustraran cada una de las situaciones identificadas.

Durante la investigación, también se estudiaron los documentos disponibles sobre los casos seleccionados (informes oficiales, artículos de investigación, artículos periodísticos publicados, etc.), para identificar sus procesos de diseño, implementación y desarrollo, con el objetivo de determinar los retos y oportunidades relacionados con cada experiencia y extraer conclusiones.

3. Véanse, por ejemplo, FaccT Conference (https://facctconference.org), el congreso de AIES (https://www.aies-conference.com) y CHI (https://chi2022.acm.org).

4. Véanse, por ejemplo, el informe *Automating Society* de AlgorithmWatch (https://automatingsociety.algorithmwatch.org) y el del OASI de Eticas Foundation (https://eticasfoundation.org/oasi/).

2. La supervisión humana, en contexto

Definición de supervisión humana

En general, el término *supervisión* se emplea en las políticas públicas a diferentes niveles, e implica transparencia institucional, responsabilidad pública o intervención en los resultados.[5] Esas implicaciones son difusas, pero ni la normativa actual ni las que se han propuesto ayudan a definirlas. En el contexto de este informe sobre políticas públicas, proponemos la siguiente definición:

> **Por *supervisión humana* se entiende, principalmente, la capacidad de actuación de los supervisores u operadores humanos en un sistema (basado en algoritmos) para mitigar los daños o errores causados por el sistema.**

Una interpretación habitual del término *supervisión humana* es la de devolver a las personas el control de un proceso que se ha automatizado. Es decir, las personas deberían estar capacitadas para intervenir lo suficiente en el sistema (y poder controlar o cambiar las decisiones de este), como medio para minimizar los riesgos.

Uno de los enfoques más populares se conoce, en el ámbito académico y los sectores técnicos, como la solución *human-in-the-loop* (HITL), lo que significa incluir a las personas en el proceso.

> **HITL es la capacidad de que las personas intervengan en cada ciclo de decisiones del sistema, con lo que vuelve a situarse al ser humano en el centro del proceso de toma de decisiones.**

Aunque el campo de la interacción persona-ordenador lleva varias décadas analizando el modelo HITL, inicialmente, HITL era una forma de integrar la intervención humana en sistemas críticos que requerían el criterio de una persona, por ejemplo, en la aviación o la robótica (Dourish 2001). Como se explicará más adelante en este documento, HITL ha pasado a ser un término más amplio a medida que las decisiones automatizadas se han ido incorporando a los servicios públicos. Esto incluye diversos ejemplos que ilustran cómo, a menudo, la intervención humana no es ni posible ni deseable, ya que puede dejar de ser una medida eficaz para limitar los daños o errores de los algoritmos.

5. Por ejemplo, Facebook ha creado recientemente un consejo de supervisión para ofrecer respuestas e incrementar la rendición de cuentas sobre determinadas decisiones tomadas por el algoritmo de la compañía (véase https://oversightboard.com).

La supervisión humana en la normativa europea

La supervisión humana en el Reglamento General de Protección de Datos (RGPD) europeo

En el artículo 22 del RGPD se intuye la noción de supervisión humana, puesto que prohíbe las decisiones individuales totalmente automatizadas (Parlamento Europeo y Consejo de la Unión Europea 2016).

Aunque esta normativa no menciona el término *supervisión humana*, sí requiere que intervenga una persona en el caso de los ADMS. El artículo 22 establece que todo ciudadano "tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado [...] que produzca efectos jurídicos en él o le afecte significativamente de modo similar".

Esta disposición incluye los sistemas empleados para elaborar perfiles personales o para proporcionar puntuaciones que determinen el acceso a prestaciones sociales, pero solo si los sistemas están totalmente automatizados. En el tercer párrafo se estipula que "[el trabajador social o representante gubernamental] adoptará las medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado, como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión" (Ibid.).

Sin embargo, el artículo 22 no es claro en cuanto a qué puede considerarse "únicamente tratamiento automatizado" y qué tipo de intervención humana es suficiente como "medida adecuada". Tampoco queda claro en qué medida los sistemas deberían estar diseñados para incluir puntos de vista discordantes y para permitir que las personas (los supervisores o revisores humanos) impugnen las decisiones automatizadas.

Es decir, el artículo se centra solamente en las decisiones tomadas por sistemas automatizados y no contempla todo el abanico de posibilidades en que las personas y las empresas u organizaciones pueden participar en sistemas de toma de decisiones (Brkan 2017).

Y es que pueden existir todo tipo de situaciones, por ejemplo, las que se dan cuando las personas recopilan los datos que se usarán para entrenar a un algoritmo, lo que implica que no es un proceso totalmente automático; cuando un sistema automatizado no toma una decisión definitiva, sino que se limita a clasificar la información sobre los ciudadanos, o cuando un asistente social tiene que aprobar la decisión final.

La propuesta de regulación pasa por alto esta gran variedad de oportunidades potenciales de supervisión humana. Por desgracia, estas oportunidades permiten que las personas —sea o no de manera intencionada— incorporen valores, sesgos y suposiciones a la toma de decisiones, por lo que se requieren una atención y una consideración mucho mayores. Si no se reconoce que la ambigüedad, la incompleción y la duda también forman parte de las decisiones humanas, no parece que incluir a personas en las decisiones asistidas por algoritmos sea una vía directa a una solución (Birhane 2021).

Por otra parte, las decisiones automatizadas en las que intervienen los datos personales de los interesados (por ejemplo, de los ciudadanos) también deben cumplir con los artículos 13 y 14 del RGPD, que exigen que los interesados tengan acceso a "información significativa sobre la lógica aplicada" (Parlamento Europeo y Consejo de la Unión Europea 2016). Porque es de sobra conocido que muchos algoritmos son cajas negras y siguen siendo complicados de explicar, incluso para los técnicos especializados (Miron 2018).

Sin embargo, puesto que el RGPD se redactó desde el punto de vista del ecosistema de los datos, estos artículos no contemplan la complejidad de los sistemas algorítmicos e ignoran las dificultades de los procesos de toma de decisiones a las que se enfrenta el personal que maneja los ADMS (los operadores humanos). Por ejemplo, la normativa actual no contempla la necesidad de explicabilidad de los ADMS ante los usuarios finales ni requiere que estos sistemas sean impugnables.

La supervisión humana en el informe sobre inteligencia artificial del grupo de expertos de alto nivel sobre la IA

Teniendo en cuenta estos artículos del RGPD, el grupo de expertos de alto nivel sobre la IA (AI HLEG) está definiendo una serie de requisitos para elaborar una regulación más concreta de la inteligencia artificial. En el 2020, el *Libro Blanco sobre la inteligencia artificial* recopiló las sugerencias del AI HLEG para establecer una regulación de la Unión Europea (Comisión Europea 2020). En la página 21 del libro blanco, dentro de la sección D(e), aparecen las "consecuencias siguientes, entre otras", de la supervisión humana:

- El resultado del sistema de IA no es efectivo hasta que una persona no lo ha revisado y validado (por ejemplo, la decisión de denegar una solicitud de prestaciones de seguridad social solo podrá adoptarla un ser humano).
- El resultado del sistema de IA es inmediatamente efectivo, pero se garantiza la intervención humana posterior para cambiar esa decisión (por ejemplo, la decisión de denegar una solicitud de tarjeta de crédito podrá tramitarse a través de un sistema de IA, pero deberá posibilitarse un examen humano posterior).
- Se realiza un seguimiento del sistema de IA mientras está en funcionamiento, y es posible intervenir en tiempo real y desactivarlo (por ejemplo, un vehículo sin conductor cuenta con un procedimiento o botón de apagado para las situaciones en las que un humano determine que el funcionamiento del vehículo no es seguro).
- En la fase de diseño, se imponen restricciones operativas al sistema de IA (por ejemplo, un vehículo sin conductor dejará de funcionar en determinadas condiciones de visibilidad reducida en las que los sensores sean menos fiables, o mantendrá una cierta distancia con el vehículo que lo preceda en una situación dada).
- Desgraciadamente, aunque estas manifestaciones son ejemplos claros de cómo concebir la supervisión humana en diferentes situaciones, el libro blanco no define los sistemas de IA de alto riesgo, pese a que la Comisión Europea publicó previamente una recomendación abordando esta cuestión.

La supervisión humana en la propuesta de Ley de Inteligencia Artificial

A principios del 2021, la Comisión Europea publicó una recomendación para la regulación de la IA, denominada *Ley de Inteligencia Artificial* o LIA (Comisión Europea 2021). En ella, se incluía la necesidad de incorporar la supervisión humana en los sistemas de IA de alto riesgo, mientras que la supervisión humana es opcional en los niveles de riesgo inferiores.

La clasificación de alto riesgo de la IA depende de lo que esté en juego, considerando si el sector y el uso previsto implican riesgos significativos. Los sistemas de IA de alto riesgo se definen en términos generales en el artículo 6 y se especifican en el anexo III de la Ley de IA. Se incluyen los ámbitos en los que se puede aplicar la automatización, como la justicia penal, los sistemas de bienestar de la infancia, la selección de recursos humanos o el uso de datos biométricos para la identificación personal.

Sin embargo, las definiciones de supervisión humana en el contexto de los sistemas de IA de alto riesgo desaparecieron de la Ley de IA, con lo que solo quedó una escueta descripción en el artículo 14 (Ibid.):

> **El objetivo de la vigilancia humana será prevenir o reducir al mínimo los riesgos para la salud, la seguridad o los derechos fundamentales que pueden surgir cuando un sistema de IA de alto riesgo se utiliza conforme a su finalidad prevista o cuando se le da un uso indebido razonablemente previsible, en particular cuando dichos riesgos persisten a pesar de aplicar otros requisitos establecidos en el presente capítulo.**

De acuerdo con la Ley de IA, los supervisores humanos deberían mostrar estas capacidades (Ibid.):

- Entender las capacidades y limitaciones del sistema de IA de alto riesgo.
- Ser conscientes de la posible tendencia a confiar ciegamente o en exceso en la información de salida generada por un sistema de IA de alto riesgo.
- Interpretar correctamente la información de salida del sistema de IA de alto riesgo.
- Decidir, en cualquier situación concreta, no utilizar el sistema de IA de alto riesgo o desestimar, invalidar o revertir la información de salida que este genere.
- Intervenir en el funcionamiento del sistema de IA de alto riesgo o interrumpir el sistema.

Puntos ciegos de la supervisión humana en la normativa de la UE

En la actualidad, el RGPD es el único reglamento vigente que requiere la supervisión humana de los sistemas automatizados de toma de decisiones, pero se espera que esto cambie radicalmente con la Ley de Inteligencia Artificial, que promete llegar más lejos y velar por que los sistemas de IA no representen un riesgo para la salud, la seguridad ni los derechos fundamentales. Aun así, se está debatiendo si la Ley de IA contempla correctamente la complejidad que supone la supervisión humana de sistemas algorítmicos.

El reglamento afronta algunos retos y puede resultar difícil de cumplir para los actores implicados, debido a la terminología que emplea, la inevitable ambigüedad de los casos que aborda y la rapidez con que evolucionan los contextos en los que se desarrolla la IA (Green 2021). Estos son algunos aspectos que explican en mayor profundidad las tensiones entre ambos:

1. La propuesta de regulación de la IA emplea palabras y términos técnicos para hacer referencia a roles concretos de los sistemas algorítmicos (como *usuario*, *proveedor*, *responsable del tratamiento* o *interesado*). **No aborda la complejidad contextual de la interacción de estos sistemas con otros sistemas y estructuras institucionales**, por lo que sienta un precedente ambiguo en cuanto a su uso e implementaciones en el mundo real.

2. **La Ley de IA da por supuesto que los ADMS son vendidos por terceros (proveedores) y adoptados por los usuarios, pero no se identifica claramente a los usuarios en el contexto de casos de uso concretos**. Por ejemplo, ¿los usuarios son los representantes de un Gobierno, los ciudadanos o el personal informático?
   Asimismo, el artículo 29, titulado "Obligaciones de los usuarios de sistemas de IA de alto riesgo", establece que el proveedor debe integrar en el sistema la supervisión humana. En caso de errores del sistema, la única medida que se estipula es que el usuario (que, una vez más, no se define claramente) debe informar al proveedor o distribuidor e interrumpir el uso del sistema.
   No se sugiere ningún otro método para mitigar esos errores. Sin embargo, se desconoce si los proveedores son organismos o empresas externas y si saben cómo funciona el sistema analógico actual y cómo se toman las decisiones.

3. Pese a los esfuerzos por proporcionar pautas para definir las situaciones al implementar la IA y los ADMS, la IA está provocando cambios en nuevos contextos, que se incrementan cada año. Los contextos evolucionan y los algoritmos se actualizan constantemente, con lo que adquieren nuevas capacidades.
   Al parecer, **los mecanismos de regulación están limitados, ya que abordan algoritmos diseñados solo para usos concretos** dentro de la toma de decisiones automatizada (véase el ya citado anexo III) y tal vez no sirvan en muchas otras situaciones. Por lo tanto, la regulación se sitúa en zonas grises que pueden ser difíciles de regular o asociar a situaciones nuevas y ambiguas, en las que estas reglas no quedan claras para las organizaciones, las empresas y los Gobiernos.

Puntos ciegos más amplios que deben tenerse en cuenta

Más concretamente, respecto a la supervisión humana, la regulación también puede acabar siendo demasiado difusa o ambigua como para garantizar eficazmente el tipo de supervisión humana "significativa" que aspira a promover. Además, en algunos casos, puede que la supervisión humana no sea la medida adecuada para mitigar el riesgo que representa la automatización. Disposiciones como las del RGPD y la Ley de IA introducen la **supervisión humana de decisiones automatizadas en zonas grises de la implementación**.

A continuación se describen varias áreas en las que puede ser problemático actuar en la práctica:

- **Dado que el RGPD prohíbe el uso exclusivo de sistemas automatizados de toma de decisiones, da la impresión de que los sistemas actuales no cuentan con ningún tipo de intervención o supervisión.** Pero, como argumentan los investigadores Reuben Binns y Michael Veale, es difícil encontrar casos reales en que las decisiones dependan únicamente de un sistema automatizado u operado sin ningún tipo de implicación o supervisión humanas (Binns y Veale 2021).
  En el contexto de los ADMS y los sistemas de alto riesgo, es habitual que participen personas en los sistemas de implementación de algoritmos, ya sea introduciendo los datos que alimentan al algoritmo o tomando la decisión final. Por ejemplo, los jueces se apoyan en evaluaciones de riesgos para tomar decisiones en las fases previas a los juicios y sentencias; los trabajadores de atención a la infancia emplean modelos predictivos para seleccionar en qué familias investigar casos de maltrato y abandono de niños, y los organismos de bienestar social usan algoritmos al determinar si se cumplen los requisitos para recibir prestaciones, un proceso que deben confirmar los agentes.
  Como se ha explicado, estas implementaciones siempre incluyen a personas en el proceso de toma de decisiones, pero estas no siempre intervienen en la decisión final. El resultado es una prohibición mal enfocada que tal vez no sea posible aplicar nunca a las situaciones reales.

- Como señalan el RGPD y la Ley de IA, para proteger valores como los derechos humanos, es fundamental aplicar métodos más "significativos" de decisión y supervisión humanas. Pero **esas intervenciones "significativas" son ambiguas y difíciles de lograr en la práctica**. Cuando hay personas supervisando estos sistemas, a menudo, esos operadores humanos no tienen suficiente formación, motivación, capacidad de rectificación, autoridad o competencias para desempeñar ningún tipo de supervisión "significativa".

- La aplicación del artículo 22 del RGPD depende de si la decisión automatizada produce efectos jurídicos en los interesados (es decir, los ciudadanos) o de si les afecta "significativamente" (es decir, si puede tener consecuencias importantes en su vida). Si produce efectos jurídicos, la decisión debe ser supervisada por una persona.
  Los efectos jurídicos se limitan a los casos en que se modifica la condición jurídica o se crean obligaciones legales (por ejemplo, la evaluación de la condición de inmigrante o el reconocimiento de un contrato legal), pero los efectos "significativos" son mucho menos precisos.

- Y, lo que es más importante, **presentar la supervisión humana como una solución para los posibles daños puede difuminar las responsabilidades.** Por un lado, aunque se dé por supuesto que un sistema cuenta con mecanismos de mitigación, las personas pueden confiar en exceso en dichos mecanismos y dejar de supervisar los sistemas adecuadamente (por ejemplo, si se limitan a certificar el resultado final). Esto abre una vía para sortear el escrutinio y las consecuencias.
  Por otro lado, la supervisión humana podría ser una excusa para desviar la atención a los operadores humanos, lo que permite a los desarrolladores y las empresas afianzar sus promesas de eficiencia y optimización, entre otras, al tiempo que dejan la responsabilidad de corregir los errores en manos de los organismos gubernamentales y los funcionarios.

En resumen, no está claro cuándo ni en qué condiciones puede la supervisión humana ser una respuesta satisfactoria a los sesgos, daños y experiencias problemáticas que se han observado al usar ADMS. En la siguiente sección, este informe aborda los ADMS situados en la intersección de varias disciplinas, para comprender en qué medida puede resultar útil y eficaz la supervisión humana.

3. Comprensión del contexto en el que se desarrolla la toma de decisiones

El rol de los operadores humanos en la toma de decisiones puede ser fundamental en ciertas situaciones, como los controles fronterizos, los sistemas de bienestar social o la justicia penal, en las que la decisión podría limitar los derechos y prestaciones de una persona o de todo un grupo social.

Al interactuar con ADMS, se espera que los operadores humanos den respuestas concretas: desde una reacción rápida en determinados contextos, como anular una alerta en el control de seguridad de un aeropuerto, hasta una respuesta profunda y elaborada en situaciones de alto riesgo, como las de los jueces en las fases de instrucción y los juicios.

Para entender el complejo contexto de la toma de decisiones, los especialistas del ámbito académico subrayan la importancia de considerar los numerosos niveles que abarca el uso de herramientas automatizadas de toma de decisiones. Comprender esta complejidad requiere estudiar cómo se comportan e interactúan las personas con las máquinas y, también, identificar el entorno organizativo, jurídico y sociocultural.

En este contexto, hay factores humanos que deben tenerse en cuenta, como la carga de trabajo de los operadores humanos, su motivación, su seguridad en sí mismos y su confianza en la herramienta automatizada. Por otra parte, también se debe atender al funcionamiento del propio sistema, que puede abarcar aspectos como su transparencia, su eficacia como herramienta, etc. (Ananny y Crawford 2018; Kemper y Kolkman 2019; Zhang et al. 2020; Lee y See 2004).

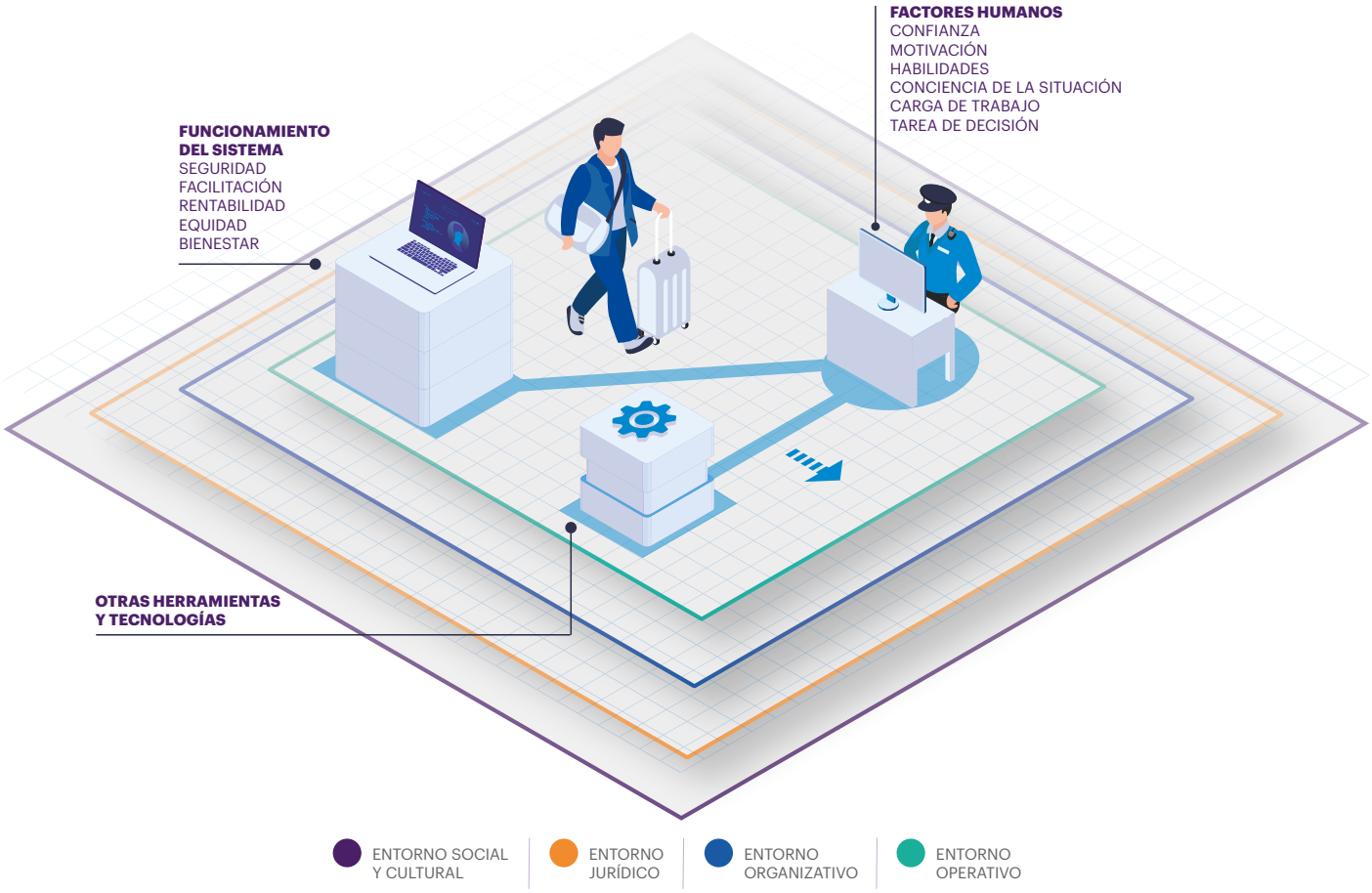
En el proyecto BODEGA,[6] un grupo de investigación que estudió el contexto de la toma de decisiones en los puntos de control fronterizo automatizados de la UE (que se analizará más adelante en la sección de estudios de caso), el marco de factores humanos (figura 1) demostró ser un modelo útil para visualizar los factores humanos que intervienen en la labor de los guardias de fronteras y su interconexión.

Este marco describe el entorno en el que operan los guardias de fronteras y define los factores que contribuyen al funcionamiento del sistema, todos los cuales pueden extrapolarse a otros contextos de toma de decisiones.

6. Para obtener más información, véase https://bodega-project.eu/.

Los entornos generales que contempla el marco incluyen los siguientes: 1) **el entorno social y cultural**, que describe los valores, las normas y la opinión pública; 2) **el entorno jurídico**, que condiciona las implicaciones legales del sistema, como el tipo de automatización o de datos que se pueden procesar; 3) **el entorno organizativo**, muy influenciado por los dos entornos anteriores, y que incluye la cultura y la estructura organizativas, y finalmente 4) **el entorno operativo**, que describe el entorno físico y espacial en el que tiene lugar la toma de decisiones.

Figura 2. **Marco de factores humanos en el control de fronteras**



Fuente de la imagen: Kulju et al. 2019.

Esta figura se ha tomado del estudio sobre los controles fronterizos e ilustra cómo se entrelazan el contexto (los entornos) de las decisiones y los factores humanos durante las tareas de toma de decisiones. Además, la interacción con el sistema y otras herramientas y tecnologías permiten o impiden que las tareas se implementen correctamente. En resumen, en esta situación, los viajeros pasan por un proceso en el que influyen varios factores.

Entrando en mayor detalle, **el entorno organizativo general** abarca las organizaciones que desarrollan el software y las que lo usan. Pueden ser los mismos actores o, como muestran más adelante los estudios de caso, las Administraciones del sector público pueden adquirir estas herramientas. Dichas organizaciones (por ejemplo, las instituciones públicas y el sector privado) influyen en gran medida sobre, entre otras cosas, qué procesos se automatizan, qué datos se usan para entrenar a los algoritmos, cómo se diseñan los sistemas y los diferentes niveles de responsabilidad que se encuentran en el sistema.

En particular, definen cómo y dónde se sitúa la supervisión humana, así como la selección de los operadores encargados de tomar esas decisiones. Este aspecto es muy importante en términos de supervisión humana, puesto que ni las personas ni los algoritmos son ajenos a los sesgos contextuales y a la exposición a decisiones sesgadas, y los sistemas deberían prever los riesgos potenciales que esto conlleva.

Respecto al entorno operativo, la interacción con dispositivos tecnológicos y sus interfaces influye en las personas que toman las decisiones. Este entorno está muy influenciado por factores humanos y afecta al rendimiento de las personas que toman las decisiones, que puede ser positivo, neutro o negativo dependiendo de la situación.

Por ejemplo, para que los operadores humanos puedan proporcionar una respuesta independiente y justificada, puede ser necesario un grado distinto de conocimientos especializados sobre el contexto de uso, desde profesionales con mucha experiencia hasta operadores con pocos conocimientos (Myers-West et al. 2019). La capacidad de un usuario de utilizar adecuadamente un ADMS puede depender de cómo proporcione el sistema la información relevante y de sus conocimientos para entender la comunicación mediante datos y elementos gráficos.

Por lo tanto, en el momento en que se toma la decisión, las personas están interactuando con algoritmos mediante la interfaz de una máquina y, al mismo tiempo, están bajo la influencia de los factores humanos descritos más arriba. Para quienes investigan la interacción persona-ordenador (HCI), es imprescindible tener en cuenta estas interacciones dentro del contexto en el que se produce la interacción (Suchman 2007).

Por ese motivo, los investigadores han tratado de crear marcos que representen todas las variables complejas que intervienen durante la interacción entre personas y máquinas. Elementos como las capacidades comunicativas, las habilidades tecnológicas, las variables personales, las actitudes o las motivaciones, entre otros, pueden determinar qué se puede esperar del comportamiento de un sistema o de la interacción con él (Cranor 2008).

“Cuando se introduce el elemento humano en el diseño de un sistema de apoyo a las decisiones, emergen niveles totalmente nuevos de cuestiones sociales y éticas, pero no siempre se reconocen como tales” (Cummings 2006).

4. La complejidad de la supervisión humana

Como se ha mencionado en la sección anterior, hay varias limitaciones que influyen en el contexto de la toma de decisiones. Estos factores, de diversas escalas, afectan a la manera en que los operadores humanos supervisan la automatización. Para la regulación, una supervisión significativa consiste en que los operadores ejerzan su poder de actuación siendo conscientes de los sesgos o las limitaciones del sistema (y de los suyos propios).

Esto implicaría que los operadores humanos pudieran evitar los daños entendiendo cuándo se equivoca un algoritmo, comprender por qué un algoritmo ha tomado una determinada decisión y considerar los sesgos potenciales del sistema. Por ello, en teoría, para que la supervisión humana sea eficaz, el diseño del sistema también debe tener en cuenta las limitaciones y los sesgos de los operadores humanos.

En esta sección se detallan varias limitaciones que pueden suponer un riesgo para que la supervisión humana sea eficaz, de modo que pueden dar lugar a una falsa sensación de seguridad. Es importante tener en cuenta los condicionantes enumerados en esta lista no exhaustiva, que más adelante resultarán útiles para analizar los estudios de caso.

1. Las personas tienen una capacidad limitada de interpretar y procesar información compleja en poco tiempo o bajo presión

Para emitir un juicio, es fundamental captar y comprender la información. Los seres humanos están dotados para emitir juicios, pero hacerlo les supone un reto (Kahneman 2011). En el contexto de los ADMS, requie re entender conceptos abstractos, leer gráficos y contextualizar información. Sin embargo, la eficiencia de las personas disminuye cuando se enfrentan a una sobrecarga de trabajo y de información (Balfe et al. 2018).

Además, las personas son menos capaces de procesar más de tres o cuatro elementos de información al mismo tiempo, mientras que los ordenadores pueden realizar cientos de cálculos simultáneamente.

Cuando la cantidad de información y de relaciones complejas entre los datos es considerable, los algoritmos se desenvuelven mejor que las personas, pero, a pesar de su capacidad de procesamiento, los algoritmos carecen de contexto (por ejemplo, no entienden el significado de la información que procesan).

Asimismo, numerosos algoritmos, como los de las redes neuronales profundas, se consideran difíciles de entender para la mente humana; son lo que se conoce como “cajas negras” (Rudin 2019). Pero no todos los tipos de algoritmos encajan en esta definición. Por ejemplo, la mayoría de las técnicas de aprendizaje automático que se usan en los ADMS, como las regresiones logísticas o los árboles de decisión, son fáciles de entender.

Todo esto significa que, según la formación y los conocimientos previos de las personas que toman la decisión, y dependiendo del tipo de algoritmo implementado, el resultado del algoritmo puede ser más difícil o más fácil de comprender y contextualizar en un tiempo limitado. Cuando las personas no han recibido la formación adecuada para entender y procesar información sensible al contexto, dicha información puede acabar siendo para ellos una fuente de confusión y desinformación, especialmente en contextos de alto riesgo social (Scurich et al. 2012; Batastini et al. 2019).

2. Las personas tienen dificultades para entender el rol del algoritmo en el proceso de toma de decisiones

Para que las personas puedan intervenir en las decisiones de un algoritmo, deben ser capaces de identificar cuándo y dónde producen errores dichos sistemas de toma de decisiones, y cómo influye el algoritmo en que se produzca el error.

Pero los algoritmos pueden desempeñar diversos roles, pueden ayudar a la persona que decide proporcionándole información para facilitar el proceso o pueden llegar a una decisión final con la información proporcionada por la persona (los distintos tipos de sistemas se describirán más detalladamente en la siguiente sección). En consecuencia, es posible que las personas no sean conscientes del rol específico que han desempeñado los algoritmos en una decisión concreta.

En un estudio que analizaba las evaluaciones de riesgos en los procesos de toma de decisiones por parte de personas, los investigadores Ben Green y Yiling Chen observaron que incorporar evaluaciones de riesgos a la predicción humana, como las que se emplean para predecir la reincidencia de los delincuentes, es una labor muy difícil que requiere más conocimientos de lo que se esperaba.

De acuerdo con ese estudio, los participantes no eran capaces de evaluar correctamente sus resultados ni los del algoritmo. El estudio descubrió que los mejores resultados se obtenían cuando los participantes seguían la evaluación de riesgos, lo cual es problemático, dados los sesgos inherentes a este tipo de evaluaciones en los sistemas de justicia penal.

Ben Green y Yiling Chen argumentan que hay pocas pruebas de que las evaluaciones de riesgos, junto con el criterio de la persona que toma la decisión, produzcan mejores decisiones, y esto confirma la falta de comprensión sobre si las evaluaciones de riesgos ayudan en los procesos de toma de decisiones y sobre cómo lo hacen (Green y Chen 2019a).

3. Las personas no suelen cuestionar las sugerencias de los algoritmos (exceso de confianza)

A menudo, las personas encargadas de tomar decisiones se limitan a certificar las sugerencias del algoritmo, sin reflexionar sobre ellas. Ben Wagner lo llama “cuasiautomatización”, y no se contempla en los marcos regulatorios vigentes ni propuestos (Wagner 2019).

Además, Ben Green y Yiling Chen han identificado varios problemas que podrían llevar a este tipo de aprobación sin reflexión previa (Green y Chen 2019a). Por ejemplo, el *sesgo de automatización*, que hace que las personas confíen demasiado en las sugerencias automáticas, o el *sesgo afirmativo*, por el que las personas tienden a estar de acuerdo con las decisiones automatizadas que coinciden con sus valores y creencias.

La supervisión humana es especialmente difícil de implementar con eficacia debido a los constantes cambios en el contexto al que deben adaptarse los algoritmos para seguir funcionando de manera correcta y relevante.

En general, los algoritmos se entrenan con determinados datos y se implementan en un sistema digital concreto. Los algoritmos pueden volver a entrenarse con nuevos datos, pero es habitual que se mantengan invariables durante semanas, meses o años, según el tipo de sistema y las necesidades.

Los algoritmos deberían actualizarse constantemente para reflejar todos los cambios sociales y culturales relevantes. Sin embargo, si los datos en los que se basan esas actualizaciones están sesgados y la supervisión humana no es capaz de reconocer esos sesgos, o se fía en exceso de las decisiones automatizadas, el resultado será que se reforzarán los prejuicios. Esta práctica puede incluso reintroducir nuevos sesgos que se habían mitigado anteriormente. De ese modo, a medida que se desarrollan nuevos algoritmos, los nuevos sesgos pasan desapercibidos.

Este fenómeno se conoce como *efecto de retroalimentación*, y puede provocar sesgos inesperados o indeseables, como las profecías autocumplidas (es decir, predicciones que acaban haciéndose realidad por el hecho de hacerlas y actuar como si fueran ciertas) o las predicciones que propician una transformación de la situación social (Barocas et al. 2019).

Por ello, interactuar con decisiones automatizadas sin un análisis adecuado puede reducir las ventajas que aporta el criterio de las personas o, en el peor de los casos, ocasionar nuevos daños.

“Debido a la complejidad inherente a los sistemas sociotécnicos, los sistemas de apoyo a las decisiones que integran niveles superiores de automatización pueden hacer que los usuarios perciban al ordenador como una autoridad legítima, que confíen menos en su propia moral y que transfieran la responsabilidad al ordenador, con lo que queda amortiguado el aspecto ético” (Cummings 2006).

4. Las personas pueden equivocarse valorando más el juicio personal que la recomendación de un algoritmo (falta de confianza)

Como se ha mencionado, la eficacia de la interacción entre personas y algoritmos depende del nivel de formación, la experiencia con el algoritmo de las personas que tomen la decisión y el ámbito profesional específico en el que se tome la decisión. Al igual que los operadores pueden confiar demasiado en las decisiones automatizadas, en el extremo opuesto del espectro puede darse una falta de confianza, es decir, que los operadores humanos no estén de acuerdo con la recomendación del algoritmo o se desvíen de ella. Esto puede deberse a diferentes motivos, dependiendo de la formación de los operadores, sus percepciones, su carga de trabajo, etc.

En algunos contextos, en los que la decisión final requiere una comprensión minuciosa del proceso por parte de una persona experimentada, es posible que los evaluadores no aprovechen todas las ventajas de la precisión que aportan las predicciones, y tomen una decisión distinta a la del algoritmo aunque las predicciones de este fueran válidas (McCallum et al. 2017).

Esto ocurre principalmente porque los evaluadores humanos no entienden el razonamiento del proceso automatizado de toma de decisiones, ya que no tienen acceso a los datos y, por lo tanto, no pueden evaluar el proceso de predicción. Las personas que toman las decisiones también pueden mostrar una *aversión a los algoritmos* y dejar de usar un algoritmo determinado después de observar un error, aunque de media el algoritmo acierte más que ellas (Dietvorst et al. 2014; Burton et al. 2020; De-Arteaga et al. 2020). Así, la persona encargada deja de confiar en el resultado del sistema.

Ocurre algo similar cuando las personas que deben tomar las decisiones tienen experiencia y, aunque se les proporcione información clara y una sugerencia correcta, tienden más a desviarse de las recomendaciones del algoritmo y a confiar en sus propios procesos cognitivos (Green y Chen 2020).

Y por último, en lugar de desconfiar del sistema, como en el caso anterior, las personas pueden desviarse de las recomendaciones del algoritmo porque sus objetivos no coincidan con el fin para el que fue optimizado el algoritmo, o porque el contexto cree incentivos para desviarse de la recomendación (Green 2020; Stevenson y Doleac 2019).

5. Definición de los distintos tipos de interacción entre personas y algoritmos

El marco de los guardias de fronteras descrito más arriba ilustra cómo la interacción persona-ordenador va más allá del mero uso de una máquina y puede observarse en diferentes escalas o entornos, por ejemplo, organizativos, jurídicos, etc. (como se representa en la figura 1).

No obstante, dado que no todos los sistemas automatizados son similares, hay ciertos aspectos complejos que se deben abordar desde el punto de vista del sector público. Para explicar en mayor detalle cómo es la automatización en el caso de los ADMS en los contextos de toma de decisiones del sector público, en este informe se hará referencia a tres tipos de automatización (véase la figura 2).

Estos tipos, definidos por los investigadores Reuben Binns y Michael Veale, son especialmente útiles para explicar cómo contribuye la automatización a la toma de decisiones por parte de las personas.

Como explican dichos investigadores, **estos tipos son versiones simplificadas de los diferentes roles que puede desempeñar la automatización y, en realidad, cada ADMS puede incluir varios de ellos**. Aun así, para que queden más claros, en este documento se analizarán por separado. Los roles son los siguientes:

- **Síntesis:** el sistema reúne datos o intervenciones humanas de una o más personas encargadas de tomar decisiones, lo que lleva a una decisión automatizada.
- **Apoyo:** el sistema proporciona información a la persona que debe tomar las decisiones y esta tiene en cuenta los "consejos" del sistema.
- **Clasificación:** el sistema procesa los casos automáticamente salvo que se marquen para que los revise una persona.

Cada tipo de automatización se puede dividir en dos períodos clave: el del proceso *ascendente* de la automatización, durante el cual se recopilan, sistematizan y procesan los datos, y el del proceso *descendente*, que tiene lugar durante las etapas de implementación y monitorización, tras el resultado automatizado.

El proceso ascendente es básicamente el de recopilar la información que se le proporciona al algoritmo, mientras que el proceso descendente abarca todo lo que sucede después de que el algoritmo haya emitido el resultado. Los tres roles mencionados permiten entender mejor cómo puede intervenir una persona en el proceso. Con un sistema de **síntesis**, por ejemplo, las personas que toman las decisiones participan en el proceso ascendente del sistema **proporcionando evaluaciones o valoraciones que se convierten en datos estructurados**.

Un ejemplo de sistema de síntesis, que examinaremos más adelante, es la evaluación de reclusos. Las evaluaciones codificadas como puntuaciones numéricas, determinadas por trabajadores sociales, se someten a un proceso automatizado que asigna puntuaciones de riesgo a los presos.

Los sistemas que proporcionan información (**apoyo**) requieren intervención humana en el proceso descendente, es decir, **las personas reciben una puntuación automática y la tienen en cuenta junto con otra información para tomar una decisión final**.

La **clasificación** produce decisiones automatizadas en las que pueden intervenir las personas tras el análisis automático inicial. Si el sistema detecta que el caso en cuestión es sospechoso o que debe revisarlo una persona, requerirá la intervención de esta en el proceso descendente.

En esta categoría se incluyen, por ejemplo, los sistemas de detección de anomalías. Udbetaling Danmark (UDK, Pagos de Dinamarca) emplea este tipo de sistema para detectar errores o fraudes en los pagos de prestaciones sociales. Por último, en el caso de los sistemas **sin intervención**, las personas no pueden participar en el proceso automatizado.

Figura 3. **Tipos de toma automatizada de decisiones e intervención humana**

Fuente de la imagen: Binns y Veale 2021.

Los estudios de caso que se examinan en la siguiente sección demuestran cómo los diferentes tipos de interacciones entre personas y algoritmos que hemos descrito se solapan entre sí en multitud de sistemas, así como sus consecuencias.

6. Estudios de caso

Tras explorar la supervisión humana, los distintos tipos de interacción entre personas y ADMS, y la complejidad que conllevan, veamos cómo es la supervisión humana en la práctica analizando tres estudios de caso. Dado que este informe se centra en la legislación europea, todos los casos seleccionados reflejan la supervisión humana de ADMS utilizados en el sector público europeo:

1. **Udbetaling Danmark (UDK, Pagos de Dinamarca):** una aplicación basada en datos para detectar errores y fraudes en los pagos de prestaciones sociales (clasificación).
2. **Control automatizado de fronteras de Frontex:** la automatización del control fronterizo en algunos Estados del espacio Schengen (clasificación).
3. **RisCanvi:** una herramienta de evaluación de riesgos que predice la reincidencia de actos violentos entre presos de Cataluña (síntesis).

Los estudios de caso elegidos se centran de manera específica en la interacción persona-máquina y en cómo se ha planeado la supervisión humana de cada herramienta. Esta sección pretende ilustrar los tipos presentados en el marco descrito por Binns y Veale, y aportan al lector una perspectiva de diferentes casos y de los retos que representa cada uno de ellos para la capacidad de decisión de las personas.

Udbetaling Danmark (UDK)

Detección de errores y fraudes en las prestaciones sociales

Contexto

Dinamarca se considera uno de los países pioneros en la transformación digital del sector público, incluida la tendencia actual de digitalizar el estado de bienestar y automatizar decisiones relacionadas con el acceso a las prestaciones. En una estrategia de gobierno digital definida en el año 2011, se concebía la digitalización como obligatoria para prestar un mejor servicio a los ciudadanos y las empresas (Deloitte y Lisbon Council 2020). Desde el 2016, el sistema de autoservicio digital danés requiere que los ciudadanos soliciten online los servicios y prestaciones públicos.

En paralelo al plan de transformación digital, el Gobierno creó UDK para centralizar los pagos que hasta entonces llevaban a cabo los diferentes municipios y facilitar su transición digital (Østergaard Madsen et al. 2022). UDK asumió la administración de los pagos relacionados con la vivienda, la discapacidad de familiares y las bajas por maternidad. La reorganización de los servicios redujo los costes administrativos y sustituyó la comunicación presencial por un sistema de contacto telefónico y digital. Se incorporaron a UDK 1.500 de los 2.000 administradores de los diferentes municipios cuyas competencias se transfirieron a esta organización. Las 500 personas restantes pasaron a encargarse de los casos de las personas que no estaban listas para la transición digital.

Además de automatizar los pagos, UDK implementó Den Fælles Dataenhed (DFD) ('Unidad de minería de datos') para detectar errores y fraudes en el sistema cruzando datos y análisis (Ibid.). El objetivo de DFD es detectar los fraudes y errores lo antes posible. De acuerdo con UDK, lo que se desea es evitar situaciones en que los beneficiarios tengan que devolver prestaciones asignadas por error (Deloitte y Lisbon Council 2020).

¿Cómo funciona?

El objeto de este estudio de caso es la aplicación empleada para detectar errores y fraudes a partir de datos de UDK. Antes de implementar DFD, la detección de posibles fraudes se realizaba tomando como base las sugerencias de los ciudadanos o la experiencia de los investigadores responsables de detectar posibles casos de fraude. Desde el 2015, DFD centra sus esfuerzos en aprender de esos casos y detectar irregularidades para evitar que el sistema conceda prestaciones a quienes no tienen derecho a recibirlas.

No obstante, en una primera etapa se centró en inspecciones concretas y en detectar irregularidades en los pagos de prestaciones ya realizados, utilizando técnicas de aprendizaje automático a fin de desarrollar un sistema preciso bajo el proyecto de DFD. El objetivo final del sistema, que se encuentra en una de sus etapas iniciales, es que el algoritmo evite los errores y los fraudes detectando irregularidades en las nuevas solicitudes y descubriendo casos complejos que suelen pasar desapercibidos a los trabajadores sociales.

¿Qué tipo de supervisión humana implica?

La compleja tarea de detectar fraudes se puede dividir entre los casos sencillos y rutinarios, y los casos más complejos (Østergaard Madsen et al. 2022). En la primera categoría, se entrena a los algoritmos para que estén pendientes de los casos habituales de fraude. El sistema señala un caso y, a continuación, en el proceso descendente, los trabajadores sociales deciden si se trata o no de un fraude.

Según un ejemplo aportado por DFD, un caso sencillo podría consistir en usar los datos para detectar si un beneficiario tiene o no derecho a recibir las prestaciones para familias monoparentales. En este caso, el sistema puede identificar "casos sospechosos" recopilando datos sobre las personas que reciben la prestación, considerando la dirección de la pareja de esa persona, el tamaño de la vivienda, etc. Los trabajadores sociales debe investigar más.

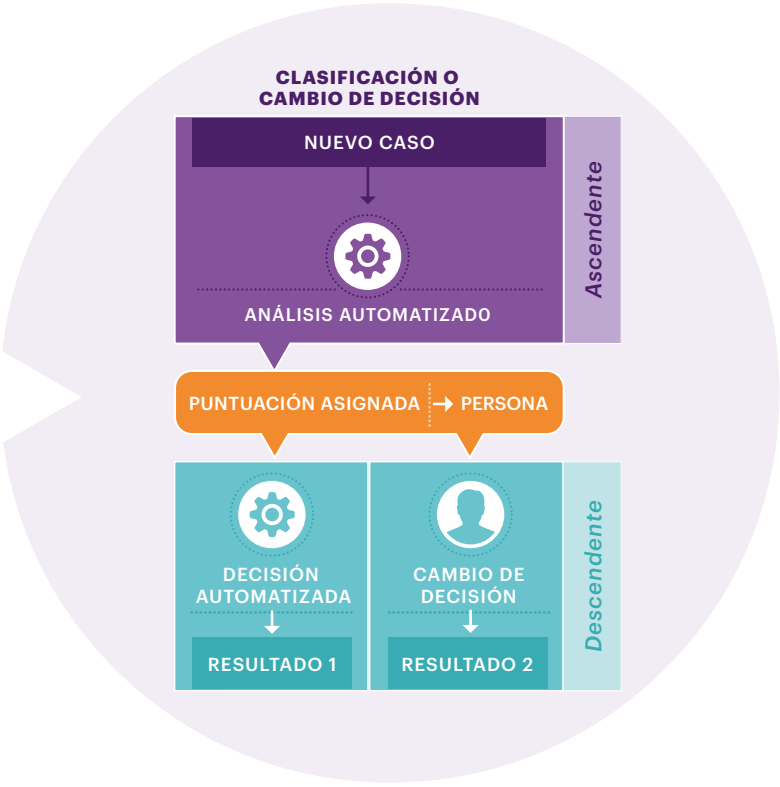
En la segunda categoría, la que comprende casos más complejos, se emplean técnicas de aprendizaje automático para detectar datos atípicos cruzando información de diferentes registros públicos (registro nacional de rentas y datos del sistema sanitario, del mercado laboral y de selecciones de personal), con el fin de reducir el número de falsos positivos y acabar detectando nuevos casos de fraude que, hasta entonces, pasaban por alto los trabajadores sociales.

Por ejemplo, cruzando datos de sanidad y de residencia, el sistema puede identificar el fraude en las bajas por enfermedad. Sin embargo, no está claro cuándo y cómo se incluye a los trabajadores sociales dentro del proceso de toma de decisiones. Los sistemas de clasificación, diseñados sin pensar en la supervisión humana, pueden dar lugar a decisiones totalmente automatizadas. En UDK hay casos que no se someten a supervisión, ya que no todos los casos que procesa son monitorizados por supervisores humanos (Eiriksson Arent 2019).

Figura 4. **Tipo de supervisión humana de Udbetaling Danmark**

Fuente de la imagen: Digital Future Society.

En este caso, consideramos que el ADMS de distribución de las prestaciones sociales puede incluir, a veces, la supervisión en el proceso descendente, ya que los municipios reciben una lista de casos prioritarios que deben analizarse más detenidamente. En todos los demás casos, todos los datos se procesan de forma automática y la decisión es también automática. No obstante, en los casos marcados para su revisión, los trabajadores sociales tienen en cuenta la decisión automatizada al tomar sus decisiones sobre qué casos son sospechosos o se han evaluado de manera errónea.



Discusión

De acuerdo con la Oficina Nacional de Auditoría de Dinamarca, la implementación de UDK ha ayudado a ahorrar 40 millones de euros anuales, al reducir el personal que trabaja a jornada completa (Østergaard Madsen et al. 2022). Los programas digitales de prestaciones sociales a menudo son objeto de críticas por dar más importancia al aumento de la eficiencia derivado de la automatización que a la larga serie de daños potenciales que pueden provocar estos sistemas, desde infringir el derecho a la privacidad hasta acusar a alguien de fraude erróneamente.

Al debatir si la implementación de la supervisión humana en UDK minimiza los daños de forma eficaz, se obtienen más preguntas que respuestas. Respecto a la asignación de prestaciones, la organización ha recibido quejas por cometer errores administrativos. En el 2019, UDK proporcionó información errónea a la Administración Tributaria danesa, y se enviaron correos electrónicos a 110.000 hogares exigiéndoles que devolvieran impuestos al organismo (Kayser Bril 2020).

La opacidad en el tratamiento de los datos

En el año 2019, la Agencia de Protección de Datos de Dinamarca (DPA) llamó la atención a UDK por recopilar datos sobre los familiares de los beneficiarios y declaró que se trataba de una infracción del RGPD. Esto ocurrió pese a que UDK afirmó que la medida estaba justificada, ya que el fin último de recopilar esos datos era detectar el fraude en las prestaciones.

Ese mismo año, Dinamarca fundó el Consejo de Ética de Datos para investigar las cuestiones éticas que conlleva compartir datos en el sector público. El caso se reabrió en el 2020, cuando la DPA cuestionó el alcance de la recopilación de datos. UDK se comprometió a eliminar todos los datos recogidos indebidamente (Eiriksson Arent 2019).

UDK no permite que los ciudadanos sepan cómo se usarán sus datos, los algoritmos que se emplean ni cómo los clasifica el sistema. Según Justitia, un laboratorio de ideas danés, esto incluye no solo a los solicitantes sino también a quienes conviven con ellos, sus cónyuges y otros miembros de la unidad familiar. Además, aunque existen dudas y críticas en torno al alcance de la recopilación de datos por parte de UDK y a cómo trata dichos datos, Justitia afirma que no es posible realizar una evaluación adecuada de la organización ni saber si cumple los requisitos que establece la ley al monitorizar a los ciudadanos (Ibid.).

La experiencia de los trabajadores sociales, en tela de juicio

Los trabajadores sociales participan en diferentes momentos del proceso de detección de errores y fraudes. Cuando se señala a un ciudadano por un posible caso de fraude o error, la información se envía a los municipios correspondientes para que filtren y analicen el caso de forma manual.

Sin embargo, el informe de Justitia sugiere que la ausencia de supervisión humana durante la etapa anterior puede ser intencionada, ya que los métodos de UDK para detectar el fraude conllevan la recopilación de una gran variedad de datos, por lo que incluir la supervisión humana en esta etapa de la automatización supondría infringir el derecho a la privacidad de los ciudadanos. Es difícil juzgar si las personas se están manteniendo fuera del proceso para resolver una mala praxis, pero es algo que se debe considerar (Ibid.).

Frontex

Control automatizado de fronteras (Automated Border Control, ABC)

Contexto

Frontex, también llamada Agencia Europea de la Guardia de Fronteras y Costas, ha señalado que el constante crecimiento del tráfico de pasajeros en las fronteras internacionales, que se espera que siga aumentando en un futuro cercano, es un reto sin precedentes. Según la agencia, los guardias de fronteras de la UE disponen de 12 segundos, de media, para evaluar al viajero que tienen delante.

Además, debido a los continuos retos políticos y administrativos, como la crisis migratoria o la amenaza del terrorismo, presentes desde hace mucho tiempo, velar por la libre circulación entre países, tanto dentro como fuera del espacio Schengen, es uno de los problemas más apremiantes de la Unión Europea.

Por ello, el escaso tiempo del que disponen los guardias de fronteras, el incremento del tráfico en los puestos de control internacionales y la gran importancia de estos controles están llevando a que se implementen tecnologías especializadas y a la adopción de sistemas de control automatizado de fronteras (Automated Border Control, ABC) (Fergusson 2014).

Siguiendo la lógica de la institución, muchos países del espacio Schengen han instalado puertas inteligentes para facilitar el tráfico de pasajeros y mejorar la eficiencia de la seguridad fronteriza. El primer país en hacerlo fue Portugal, en el año 2008, y aunque en los controles de fronteras de la UE se emplea un nivel de automatización bajo que depende en gran medida de los operadores, la demanda está aumentando. En el 2019, según los registros, había puertas ABC funcionando en más de 50 aeropuertos (Noori 2022).

La Comisión Europea, reflejando el interés por automatizar los controles de fronteras, inició un proyecto piloto denominado ABC4EU (Control automatizado de fronteras para Europa), que se desarrolló del 2014 al 2016 con el objetivo de armonizar las puertas ABC que procesan la entrada en la UE de ciudadanos de países extracomunitarios. La armonización consistió, en términos generales, en actualizar los sistemas actuales de puertas ABC para que fueran más flexibles y para incentivar su uso. Esta iniciativa piloto también pretendía evaluar el impacto de las puertas ABC, valorando el proceso de automatización e identificando posibles obstáculos (Comisión Europea 2022).

¿Cómo funciona?

Actualmente, las puertas ABC son puertas electrónicas semiautomáticas: están equipadas con lectores de documentos, dos barreras físicas y escáneres biométricos. El proceso ascendente es automático, ya que los viajeros con pasaporte electrónico (o pasaporte-e) pueden pasar por un portal ABC en el que las puertas electrónicas escanean sus documentos y consultan las bases de datos.

El sistema escanea el rostro del viajero y lo compara con los datos biométricos guardados en el pasaporte electrónico. El objetivo de este procedimiento, que antes se realizaba manualmente, es evitar tareas repetitivas a los guardias de fronteras y permitir que se centren en inspeccionar e interrogar a las personas señaladas por el sistema. Por ello, la supervisión humana se concentra en el proceso descendente.

¿Qué tipo de supervisión humana implica?

En comparación con los tipos de supervisión humana mencionados anteriormente, el sistema ABC presenta un mecanismo de cambio de decisión en el que los operadores solo actúan cuando el sistema deja de funcionar o detecta alguna anomalía.

Durante el proyecto de investigación BODEGA, financiado por la UE, se entrevistó a varios guardias de fronteras para analizar cómo afectaba el sistema ABC a su trabajo. Durante el estudio, los guardias de fronteras expresaron su desconfianza en la capacidad del sistema para reemplazarlos. Debido a la automatización, el rol de los guardias pasó de ser activo (controlar a los viajeros) a pasivo (supervisar la comprobación automática): tenían que reiniciar el sistema cuando surgían problemas de hardware o revisar los documentos cuando no se detectaban correctamente.

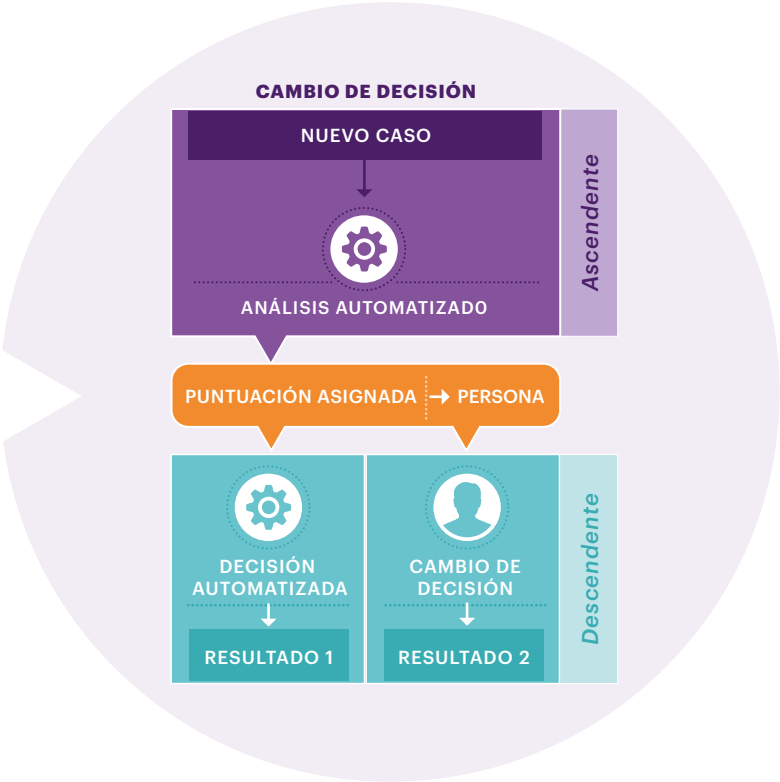
Discusión

Como se ha mencionado antes, la presión por que los ABC reemplacen en parte a los guardias de fronteras se debe al aumento de viajeros, y la implementación debería permitir que los guardias se ocupen de tareas más complejas. Sin embargo, la automatización se considera también una forma de reducir los sesgos y errores en el control de fronteras.

Figura 5. **Tipo de supervisión humana de ABC**

Fuente de la imagen: Digital Future Society.

La automatización se centra en el proceso ascendente, al escanear los pasaportes y los rostros de los viajeros para compararlos con la información de las bases de datos. En el proceso descendente, los casos en que el sistema señala a un pasajero como sospechoso o los detectados por los supervisores humanos se tratan de forma manual. En todos los demás casos, es un proceso automatizado el que toma la decisión.



En consecuencia, los expertos defienden que las fronteras inteligentes provocan un alto grado de desconfianza hacia los guardias de fronteras, al cuestionar sus competencias y capacidad de verificar la identidad, y en la práctica los convierten en "problemas" de seguridad (Noori 2020). No obstante, a menudo los guardias lo ven de otro modo y no confían en que el sistema automatizado pueda hacer correctamente su trabajo.

Por estos motivos, la implementación de puertas ABC ha ocasionado una desconfianza generalizada hacia el sistema automatizado y, de paso, en el rol de los propios guardias de fronteras. El hecho de tener que solucionar los fallos del sistema ha hecho que los guardias desconfíen aún más de ABC y lo consideren incapaz de sustituirlos en su función. La actitud de los guardias, que provoca falta de seguridad y confianza hacia el sistema, obstaculiza la eficacia global del control automatizado de fronteras (Ibid.). En esta situación, la supervisión solo es significativa cuando se detecta un error y se opta por un proceso manual. Dado que los errores son difíciles de entender y de resolver, siguen existiendo muchos puntos ciegos en cuanto a la seguridad y las políticas fronterizas.

Además, Frontex prevé que las futuras puertas ABC estén diseñadas para funcionar de manera intuitiva y requerir muy pocos conocimientos técnicos por parte de los guardias de fronteras (European Border and Coast Guard Agency 2021). En la actualidad, los guardias que han trabajado con puertas electrónicas no tienen acceso a información específica sobre cómo funciona el sistema y, más concretamente, a qué se deben sus fallos. Por ello, uno de los riesgos es la aversión a los algoritmos, que hace que los guardias de fronteras se enfrenten por su cuenta a los errores del sistema de forma inadecuada. Para mitigar este riesgo, es necesario un cierto nivel de formación o un personal que entienda bien las técnicas de IA (Ibid.).

RisCanvi

Evaluación del riesgo delictivo

Contexto

Un instrumento de evaluación de riesgos (RAI) es un tipo de herramienta algorítmica cuyo objetivo es predecir el riesgo de mala conducta de los acusados en el futuro, y suele emplearse para las decisiones judiciales previas al juicio. Las evaluaciones, en los sistemas de justicia penal, se basaban en el criterio de los profesionales, hasta que se generalizó el uso de RAI en los años setenta.

Dado que los RAI emiten predicciones estructuradas y basadas en pruebas, se introdujeron para reducir la necesidad de tomar decisiones y mejorar la objetividad. No obstante, la imparcialidad que prometen ofrecer estas evaluaciones sigue siendo un tema de debate (Heilbrun et al. 1999). Las innovaciones más recientes, como la incorporación de algoritmos informáticos en la última década, han intensificado esta preocupación, puesto que, según diversos estudios, los RAI basados en algoritmos pueden funcionar mejor que la predicción humana (Tan et al. 2018; Green y Chen 2019b).

Preocupa la cuestión de si estas máquinas ofrecen el mismo nivel de precisión y equidad que los operadores humanos, pero este tipo de herramientas también son controvertidas porque en ellas se han observado sesgos de raza y género, como en el caso del sistema de evaluación de infractores Offender Assessment System (OASys) del Reino Unido (Angwin et al. 2016). OASys, comparable a RisCanvi, generaba distintas predicciones en función de la raza, el género y la edad (Big Brother Watch 2020).

RisCanvi (derivado de las palabras “riesgo” y “cambio” en catalán) es un RAI que se emplea en Cataluña. Esta herramienta, creada en el 2009 para ayudar a los criminólogos y los trabajadores sociales a mejorar el tratamiento de los presos, se basa en varios análisis clínicos. El Departamento de Justicia de la Generalitat de Cataluña encargó la creación de RisCanvi al Grup d’Estudis Avançats en Violència (‘Grupo de estudios avanzados sobre violencia’) de la Universidad de Barcelona.

¿Cómo funciona?

RisCanvi comprende 43 factores de riesgo evaluados por profesionales, a partir de los registros y las entrevistas personales del recluso. Dichos factores de riesgo se pueden relacionar con la actitud y la personalidad del recluso y su historia clínica y personal, así como su respuesta a los tratamientos. Entre los factores están el abuso de drogas y alcohol, los antecedentes de enfermedades mentales y si el interno ha sido víctima de violencia.

El algoritmo utiliza estos factores para evaluar el riesgo de que se produzcan cinco resultados: 1) violencia autodirigida, 2) violencia dirigida a otros internos o al personal, 3) reincidencia, 4) reincidencia con violencia y 5) incumplimiento de la libertad condicional.

Un equipo multidisciplinar de profesionales recoge datos sobre cada factor junto con la historia clínica, observaciones y entrevistas. A continuación, esos profesionales introducen la información en la herramienta, que asigna al recluso una puntuación de riesgo. El resultado del algoritmo es solamente una clasificación con tres niveles, según la cual el riesgo puede ser bajo, medio o alto.

El equipo valora la evaluación final para determinar el tipo de tratamiento que recibirá el interno. La evaluación de los internos se realiza, como mínimo, cada seis meses. Los niveles indicados por las predicciones también se incluyen en los informes que se envían a los fiscales y los jueces para que decidan o no aplicar la libertad condicional.

¿Qué tipo de supervisión humana implica?

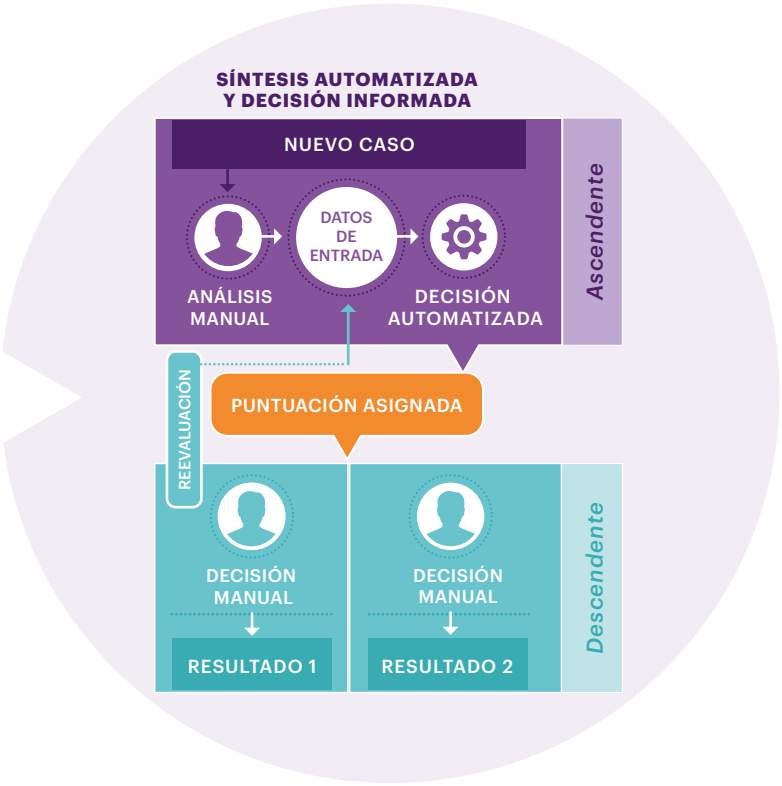
Esta herramienta incorpora la supervisión humana en dos momentos del proceso, tanto en la etapa ascendente como en la descendente. En primer lugar, los trabajadores sociales recopilan y procesan las pruebas para proporcionar datos al algoritmo. Luego, un profesional con experiencia supervisa una discusión de cada factor entre distintos profesionales y valida o ajusta el resultado de su discusión.

La información emitida por el algoritmo y las recomendaciones profesionales tienen dos consecuencias, Por un lado, el informe se envía a los fiscales y los jueces para informarlos del tipo de sentencia que deben dictar. Al mismo tiempo, los trabajadores sociales y los psicólogos usan esta herramienta para realizar un seguimiento de los tratamientos correccionales de los internos y reevaluar su estado. Esto significa que la decisión no es exactamente una decisión, por lo que no está totalmente automatizada, ya que los profesionales interactúan con información sintetizada antes de tomar una decisión informada.

Figura 6. **Tipo de supervisión humana de RisCanvi**

Fuente de la imagen: Digital Future Society.

En el proceso ascendente, los datos se introducen de forma manual durante una primera evaluación realizada por los trabajadores sociales. Después, un análisis automatizado asigna una puntuación. Hay dos tipos de usuarios, los trabajadores o asistentes sociales y los fiscales o jueces, que emplean con distintos fines dicha puntuación junto con otra información. Los asistentes sociales pueden reevaluar a los presos y cambiar la puntuación si lo consideran necesario. En ambos casos, la decisión final siempre está sujeta al criterio de una persona.



Discusión

RisCanvi ha evolucionado considerablemente desde que se creó, puesto que en un principio se planteó como una guía para la administración de prisiones, más que para dictar sentencias penales. Sin embargo, en la actualidad se tiene en cuenta para ambos procesos de toma de decisiones, lo que ha llevado a muchos a cuestionar cómo afecta el algoritmo a la toma de decisiones.

Como se ha mencionado, los RAI supuestamente deberían facilitar un debate sobre el impacto del nivel de riesgo, pero se ha observado que los jueces y fiscales, a menudo, pasan por alto las recomendaciones del informe y basan sus decisiones únicamente en las propuestas del algoritmo (Saura y Aragó 2021). Esta aprobación sin reflexión previa no se consideró un riesgo al crear el algoritmo, ya que los diseñadores esperaban que los casos se analizaran en profundidad.

Los equipos que introducen y analizan los datos reciben formación sobre el uso de RisCanvi, por lo que deberían tener suficiente experiencia con la herramienta como para saber lo que pueden esperar del algoritmo. Estos conocimientos pueden contribuir a que ignoren las predicciones en caso de que se obtengan resultados inesperados. Además, los trabajadores sociales deberían usar la herramienta para mejorar sus decisiones, contrastando sus conclusiones con las del algoritmo. Pese a esto, se ha cuestionado el escaso porcentaje de casos en que estos equipos multidisciplinares modifican el resultado del algoritmo.
Por otra parte, el sistema se ha sometido a pocas auditorías externas, lo que suscita preocupación por la falta de transparencia y los posibles sesgos que podrían afectar a algunas poblaciones infrarrepresentadas (Planas Bou 2021). Los sesgos pueden provenir de los datos introducidos, que se basan principalmente en las entrevistas de los trabajadores con los internos.

Si esta información inicial está sesgada, puede crear un efecto de retroalimentación: los datos sesgados del sistema se usan en las futuras actualizaciones del algoritmo, lo que refuerza dichos sesgos. Es lo que ha ocurrido en sistemas RAI como COMPAS (véase el caso de OASys citado más arriba), que exponen a grupos sociales enteros a la discriminación algorítmica, a raíz del uso de datos históricos.

Una lección positiva es que la interacción personal entre profesionales es fundamental para que el sistema se use correctamente. Aunque los jueces pueden hacer caso omiso de los informes de los trabajadores sociales y favorecer el criterio del algoritmo, los trabajadores consideran que es positivo transmitir su perspectiva profesional a otros profesionales. A este respecto, la confianza es esencial para mejorar la colaboración entre personas y algoritmos, al permitir un intercambio más estructurado de información y la paridad entre diferentes roles.

Principales lecciones

En la sección anterior se han expuesto tres estudios de caso que ilustran la gran complejidad de las situaciones en que se usan los ADMS y, lo que es más importante, se subraya la complejidad que hay detrás de la automatización de procesos.

En conjunto, los estudios de caso han demostrado cómo, en teoría, utilizar un algoritmo para llevar a cabo tareas habituales podría evitar que los operadores humanos tengan que realizar tareas mentales intensas que suelen resultarles difíciles, y permitir que investiguen más los casos que requieren mayor atención. Al mismo tiempo, los sistemas automatizados pueden aumentar la capacidad humana de identificar y corregir errores, y evitan así el sesgo de automatización (exceso de confianza) o la aversión a los algoritmos (falta de confianza) (De-Arteaga et al. 2020).

Supongamos que todos los algoritmos fueran explicables y transparentes. En ese caso, las personas podrían considerar e interpretar la información proporcionada y actuar en consecuencia; pero esto no suele ocurrir. En la literatura sobre comunicación de riesgos, queda claro que la información contextual y los conocimientos especializados son esenciales para tomar decisiones informadas (Heilbrun et al. 1999). Sin embargo, pueden surgir muchas situaciones inesperadas para las cuales el algoritmo no ha sido entrenado, lo que provoca errores, y las personas deberían estar preparadas para actuar en esas situaciones.

Una solución podría ser un proceso colaborativo entre personas, lo que podría mejorar su percepción y hacer que concuerde más con la de los algoritmos, como en la interacción entre profesionales que se describe en el caso de RisCanvi (Van Berkel et al. 2019). Pero los procesos colaborativos requieren más tiempo y recursos, y podrían reducir la eficiencia que prometen los ADMS. Aun así, podría considerarse esta opción para las situaciones en que las consecuencias representen un riesgo alto para la sociedad y el tiempo no sea una variable determinante, como en el caso del sistema ABC. Los enfoques colaborativos son una oportunidad que no se suele tener en cuenta en otros casos en los que podrían ayudar a reducir los sesgos y los errores.

Sigue abierto el debate sobre si la automatización altera la responsabilidad de las personas por una decisión final. Gran parte de la normativa y las presuposiciones provienen de la idea de que la supervisión humana permitiría atribuir a alguien la responsabilidad sobre los resultados (Wagner 2019). Los casos de ABC y UDK demuestran que la mayoría de los sistemas no están diseñados pensando en la responsabilidad humana: por ejemplo, si las personas solo pueden mitigar los errores o si solo son responsables de la decisión final o de proporcionar información nueva al algoritmo. Esto disminuye el poder de decisión de los supervisores humanos. La relación entre la preocupación por la protección y la privacidad de los datos, y el proceso de entrenamiento del algoritmo, en el caso de UDK, muestra que la manera de aplicar procesos de supervisión sigue siendo muy imperfecta.

Además, cuando un sistema algorítmico se diseña para incluir la supervisión, los ADMS pueden usarse de un modo distinto del previsto. Como se ha observado en el caso de RisCanvi, hay diferentes "usuarios" del sistema —los jueces, por ejemplo— que se apoyan en el resultado automatizado como parte de sus procesos de toma de decisiones, lo que puede ser más complejo de analizar. Dichos sistemas deberían articularse con los distintos contextos mencionados en la figura 1, incluyendo los factores humanos, los valores y las prácticas organizativos y las presunciones sociales y culturales. Los sistemas también deberían tener en cuenta los diferentes grados de responsabilidad, tanto del sistema como de los operadores humanos y las instituciones responsables de su implementación. Deberían poder actuar de acuerdo con sus respectivos grados de responsabilidad.

Otra tendencia identificada en los estudios de caso es la de tratar al personal como un eslabón débil del sistema o disminuir su capacidad de actuación. En el caso ABC, tratar de este modo al personal pone en riesgo todo el sistema, dado que elimina su capacidad de actuar. En el caso de UDK, el personal queda excluido del procedimiento y se usa como una medida de seguridad secundaria cuando el daño ya está hecho. Si los operadores actuaran con suficiente información, formación y experiencia, esto no solo mitigaría las consecuencias de la falta de coherencia de los algoritmos, sino que también aumentaría la confianza y mejoraría la experiencia de los ciudadanos (Chouldechova 2017).

En los sistemas de IA de alto riesgo, debería pensarse detenidamente en el rol de los usuarios (o los supervisores), que tendría que definirse junto con las capacidades y funciones del algoritmo, y teniendo en cuenta los conocimientos especializados necesarios. Mirar más allá de la supervisión humana es considerar que las personas ejercen un rol activo dentro de la complejidad del sistema.

7. Recomendaciones sobre políticas

En la mayoría de las situaciones, los algoritmos se suelen implementar para ayudar a las personas que toman decisiones, más que para que actúen de manera autónoma (Green y Chen 2020). El desarrollo de un mecanismo de supervisión humana como solución para los sesgos de los algoritmos y los daños producidos por sus incoherencias es algo que se debe pensar detenidamente.

En muchas decisiones gubernamentales, es necesario combinar predicciones precisas con otros objetivos sociales (por ejemplo, una distribución equitativa de los recursos, un tratamiento justo de los ciudadanos, etc.). Las siguientes recomendaciones generales de políticas abordan los diferentes sacrificios y aspectos complejos que se explican en este informe, pero algunas de estas recomendaciones pueden aplicarse en mayor o menor medida dependiendo del caso o el contexto particular.

Definir el nivel mínimo de implicación de las personas

La primera recomendación para implementar algoritmos con supervisión humana es pensar de qué manera ayudará la supervisión a mitigar errores de forma significativa. El RGPD y la Ley de IA destacan esta característica, pero no suele estar claro cuándo es significativa una contribución (Green 2021). La implicación de las personas puede ser superficial o transmitir una falsa sensación de seguridad. Una tarea importante es definir el sistema y analizar los sacrificios que se deben hacer en relación con las oportunidades y los retos que conlleva incluir la supervisión humana en el sistema. Por ejemplo, el diseño del proyecto ABC no incluyó a los guardias de fronteras en la implementación del sistema y generó desconfianza entre ellos, con lo que se ofrece solamente una falsa sensación de seguridad. Si una implementación no aprovecha el valor de los recursos humanos, será ineficiente y una fuente de defectos en el proceso.

Los sistemas automatizados no se deben adoptar solamente para justificar reducciones de personal o facilitar el uso de personal no cualificado. Esto puede perjudicar al funcionamiento del sistema y hacer que los supervisores se limiten a aprobar sin reflexionar las decisiones del algoritmo. Los operadores humanos deben tener potestad para impugnar y mitigar las posibles amenazas, además de poder estar de acuerdo con el sistema. Entender el gran valor de estas acciones mejoraría la eficacia del personal (Almada 2019). En los modelos dobles de toma de decisiones, “las acciones se evalúan según el grado en que ayudan a los participantes a generar información válida y útil” (Argyris 1976, p. 368).

Tener cuidado con la dependencia del contexto en que se sitúa la automatización

De acuerdo con la Ley de IA, la posibilidad de automatizar fácilmente las decisiones depende del contexto, o de si la automatización supone un riesgo alto y requiere medidas de seguridad como la supervisión humana. Pero las interacciones con los algoritmos y las estructuras de datos no se producen de forma aislada (Seaver 2019). Los sistemas algorítmicos, al igual que cualquier otro sistema tecnológico, están interconectados con otros objetos, procesos y personas (Jung et al. 2008). Debe llevarse a cabo una evaluación correcta del contexto, la infraestructura y las personas implicadas en todas las implementaciones con algoritmos que determinen una toma de decisiones, o que medien en ella, como se muestra en la figura 1.

No basta, pues, con una evaluación previa y una valoración del impacto. Para analizar los efectos que pueden tener estos algoritmos, también se deben prescribir un plan de gestión de riesgos y un proceso de aseguramiento de la calidad. Además, todas las pruebas y evaluaciones se deben realizar tanto en laboratorio (sin posibilidad de riesgos externos) como en contexto (en el que pueden detectarse errores y fallos de funcionamiento del sistema).

Los resultados de ambas evaluaciones pueden llevar a que se interrumpa el desarrollo de los algoritmos, se cambie el tipo de supervisión humana (a otro grado de implicación o a otro momento en el que intervenir) o se considere que la interacción entre ambas partes no es posible y se debe modificar todo el sistema. Sin un análisis adecuado del contexto, puede aumentar el riesgo del sistema y es posible que no se prevean los errores que podrían aparecer tras implementarlo.

Optar por sistemas abiertos y no cerrados

En el mundo del software, los sistemas cerrados representan una propiedad intelectual rigurosamente protegida y no permiten ningún tipo de intercambio abierto ni de desarrollo cooperativo de código. En cambio, el software de código abierto permite que los desarrolladores externos revisen el código y sugieran cambios para mejorar el software. Muchos de los problemas que se describen en este informe se deben al hecho de que los algoritmos forman parte de sistemas cerrados. Al igual que en el caso ABC, los usuarios no pueden ver cómo funcionan los algoritmos ni cómo se relacionan con las demás tecnologías que usan para cumplir sus objetivos. Dado que los sistemas están interconectados, las soluciones basadas en algoritmos también deben ser transparentes para interactuar mejor.

Los sistemas abiertos pueden someterse a pruebas y explicarse, por lo que los desarrolladores y los usuarios pueden identificar correctamente los errores. Esto puede contribuir a las estrategias para mitigar dichos errores y a que se avise a los usuarios cuando se produzcan situaciones inesperadas, lo que mejoraría la fiabilidad y la confianza en los sistemas. Además, los sistemas abiertos ofrecen más oportunidades de mantener adecuadamente las tecnologías, lo cual reduce las fricciones y los costes que conllevan. Por el contrario, los sistemas cerrados no son explicables y no se pueden someter a pruebas ni adaptar. Impiden que los usuarios señalen los posibles problemas y provocan aversión y falta de confianza en los algoritmos.

Definir un plan de gobernanza y diversos grados de responsabilidad

Todos los sistemas deben contar con un plan de gobernanza para definir los problemas que abordarán y gestionar la forma en que se tomarán las decisiones. Cada vez es más habitual que los ADMS implementados en el sector público tengan la obligación de llevar asociado un registro público. Por ello, dicho plan de gobernanza debería incluir también los datos, el código y las interfaces a los que pueden acceder y que pueden inspeccionar las diferentes comunidades. Esto favorecería la confianza de los ciudadanos y el consenso sobre cómo abordar las incoherencias en los efectos de ciertas valoraciones de casos (Van Berkel et al. 2019). Dado que quienes implementan los ADMS son servicios públicos que desean aumentar la eficiencia y la objetividad, es deseable prestar atención al reparto de las responsabilidades. Como ya se ha mencionado, parte de la responsabilidad se ha traspasado a los proveedores de tecnología, y reintroducir la supervisión humana sin definir las responsabilidades supondría un problema.

En los estudios de caso también se ha observado que los algoritmos, a veces, parecen tener más autoridad que las personas que los supervisan, lo que abre la puerta a que los supervisores aprueben los resultados sin reflexionar. Así, los usuarios no asumirían la responsabilidad de sus decisiones. El efecto de esta "amortiguación moral" (distanciarse de la decisión) podría ser problemático para los objetivos de rendición de cuentas que deben cumplir las instituciones públicas. Por el contrario, si el proceso es transparente y ayuda a que se identifiquen y divulguen las prácticas estándar, incluyendo la posibilidad de que la sociedad civil las inspeccione, esto permitiría introducir la supervisión humana de manera eficaz con suficiente apoyo y autoridad.

Ofrecer formación y promover la puesta en común de conocimientos entre los desarrolladores y los operadores

La normativa debería exigir que se proporcionen una formación y unos recursos adecuados al personal que opere y supervise los sistemas de IA. Además, deberían registrarse correctamente tanto su experiencia previa como los conocimientos adquiridos, lo cual ayudaría a distribuir recomendaciones y estrategias de mitigación, además de reforzar la fiabilidad y la confianza entre los operadores humanos. Por otra parte, poner en común lo que saben los desarrolladores sobre el sistema y lo que conocen los operadores sobre la labor que se debe desempeñar contribuye a catalogar los errores, lo cual permite actualizar continuamente el sistema y el algoritmo. Si se niega el acceso a estos conocimientos, se estará fomentando una organización menos dinámica, con más retrasos en la resolución de problemas y fallos de comunicación entre los equipos.

Definir un procedimiento de denuncia de irregularidades

Los sistemas persona-máquina pueden ser totalmente fiables, pero siempre hay alguna posibilidad de que se provoquen daños y de que pasen inadvertidos para las personas implicadas. En ciertos casos, las quejas oficiales sobre fallos, malas praxis o decisiones sesgadas pueden poner en riesgo el trabajo y la vida de los supervisores humanos. Dado que, a menudo, los algoritmos se consideran más objetivos y fiables que las decisiones humanas, los supervisores pueden sentirse vulnerables a la hora de tomar una decisión que contradiga al algoritmo. Es fundamental diseñar mecanismos de denuncia de irregularidades y otras medidas de protección contra posibles represalias que amparen a los trabajadores cuando estos opten por cambiar la decisión de un algoritmo, impugnar cualquier decisión automatizada o denunciar un fallo del sistema. De lo contrario, es posible que prefieran aceptar sin más las decisiones del algoritmo para evitar problemas con las autoridades y conservar su puesto de trabajo.

8. Conclusión

Este informe de políticas ha explorado y ha puesto de relieve la complejidad que conlleva implementar la supervisión humana, para facilitar el uso gubernamental de ADMS en diferentes situaciones. Lamentablemente, la consideración de la supervisión humana en la normativa vigente y en las propuestas de regulación publicadas hasta ahora es insuficiente. Las recomendaciones sobre políticas que se describen en este informe arrojan algo de luz sobre los sacrificios, las oportunidades y los retos asociados a esta cuestión en la actualidad.

Las Administraciones públicas y los Gobiernos deben pensar en el nivel de supervisión humana que es preciso implementar en los sistemas de IA de alto riesgo, en función de cada contexto particular y de las capacidades de los sistemas y del personal. Asimismo, los operadores deben recibir formación para comprender los sacrificios que implica usar estos sistemas, y se les debe ofrecer la oportunidad de aprender y entender cómo funcionarán los sistemas, además de participar activamente en su proceso de diseño.

Los tres casos presentados reflejan sistemas de algoritmos con marcos organizativos complejos. Las operaciones precisas y las posibles ramificaciones de los efectos del algoritmo pueden ser casi inextricables, de modo que puede ser necesario implementar mecanismos adecuados de transparencia y rendición de cuentas. Además, estos sistemas son difíciles de entender incluso para los expertos y los profesionales. Por ello, debe elaborarse un plan de gobernanza que permita inspeccionar el sistema y denunciar con seguridad cualquier error o daño causado durante su uso.

Implementar mecanismos de supervisión humana en los ADMS sin que su único objetivo sea mitigar los errores puede aportar numerosos beneficios a la sociedad. Las personas pueden contribuir a que los sistemas sean más seguros y se ajusten más a la normativa, mientras que las capacidades de los sistemas informáticos y la automatización ofrecen el potencial de favorecer enormemente a la sociedad. La supervisión humana no es sencilla de implementar y se debe introducir con mucho cuidado y atención. Para evitar posibles daños, es necesario tomar medidas como las recomendadas en este informe.

Referencias

Almada, M. (2019). Human Intervention in Automated Decision-Making: Toward the Construction of Contestable Systems. Forthcoming, 17th International conference on Artificial Intelligence and Law (ICAIL 2019), pp. 2–11. [PDF] Disponible en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3264189 (Consultado: 13-9-2022)

Ananny, M. y Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media and Society, 20(3), pp. 973–989. [online] Disponible en: https://journals.sagepub.com/doi/10.1177/1461444816676645 (Consultado: 13-9-2022)

Angwin, J., Larson, J., Mattu, S. y Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. [online] Disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Consultado: 13-9-2022)

Argyris, C. (1976). Single-loop and Double-Loop Models in Research on Decision Making. Administrative Science Quarterly, 21, pp. 363–377. [PDF] Disponible en: https://www.jstor.org/stable/2391848 (Consultado: 13-9-2022)

Balfe, N., Sharples, S. y Wilson, J. R. (2018). Understanding Is Key: An Analysis of Factors Pertaining to Trust in a Real-World Automation System. Human Factors, 60(4), pp. 477–495. [online] Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5958411/ (Consultado: 13-9-2022)

Barocas, S., Hardt, M. y Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. [Prepublicación]. [online] Disponible en: https://fairmlbook.org (Consultado: 13-9-2022)

Batastini, A. B., Hoeffner, C. E., Vitacco, M. J., Morgan, R. D., Coaker, L. C. y Lester, M. E. (2019). Does the Format of the Message Affect What Is Heard? A Two-Part Study on the Communication of Violence Risk Assessment Data. Journal of Forensic Psychology Research and Practice, 19(1), pp. 44–71. [online] Disponible en: https://www.tandfonline.com/doi/abs/10.1080/24732850.2018.1538474 (Consultado: 13-9-2022)

Big Brother Watch. (2020). Big Brother Watch briefing on Algorithmic Decision-Making in the Criminal Justice System. [PDF] Disponible en: https://bigbrotherwatch.org.uk/wp-content/uploads/2020/02/Big-Brother-Watch-Briefing-on-Algorithmic-Decision-Making-in-the-Criminal-Justice-System-February-2020.pdf (Consultado: 22-9-2022)

Binns, R. y Veale, M. (2021). Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR. International Data Privacy Law, 11(4), pp. 319–332. [PDF] Disponible en: https://academic.oup.com/idpl/article/11/4/319/6403925 (Consultado: 13-9-2022)

Birhane, A. (2021). The Impossibility of Automating Ambiguity. Artificial Life, 27(1), pp. 44–61. [online] Disponible en: https://direct.mit.edu/artl/article-abstract/27/1/44/101872/The-Impossibility-of-Automating-Ambiguity (Consultado: 13-9-2022)

Brkan, M. (2017). AI-supported decision-making under the general data protection regulation. Proceedings of the International Conference on Artificial Intelligence and Law, pp. 3–8. [online] Disponible en: https://dl.acm.org/doi/10.1145/3086512.3086513 (Consultado: 13-9-2022)

Burton, J. W., Stein, M. K. y Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making, 33(2), pp. 220–239. [PDF] Disponible en: https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.2155 (Consultado: 13-9-2022)

Campolo, A. y Crawford, K. (2020). Enchanted Determinism: Power without Responsibility in Artificial Intelligence. Engaging Science, Technology, and Society, 6, p. 1. [online] Disponible en: https://www.researchgate.net/publication/338486570_Enchanted_Determinism_Power_without_Responsibility_in_Artificial_Intelligence (Consultado: 13-9-2022)

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), pp. 153–163. [PDF] Disponible en: https://arxiv.org/pdf/1610.07524.pdf (Consultado: 13-9-2022)

Comisión Europea. (2020). Libro blanco sobre la inteligencia artificial - un enfoque europeo orientado a la excelencia y la confianza. [PDF] Disponible en: https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf (Consultado: 13-9-2022)

Comisión Europea. (2021). Propuesta de reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de Inteligencia Artificial) y se modifican determinados actos legislativos de la Unión COM/2021/206 final. Comisión Europea, 0106, pp. 1–108. [PDF] Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52021PC0206 (Consultado: 13-9-2022)

Comisión Europea. (2022). ABC Gates for Europe. Cordis EU research results.[online] Disponible en: https://cordis.europa.eu/project/id/312797 (Consultado 22-9-2022)

Cranor, L. F. (2008). A framework for reasoning about the human in the loop. Proceedings of the 1st Conference on Usability, Psychology, and Security, p. 15. [online] Disponible en: https://dl.acm.org/doi/10.5555/1387649.1387650 (Consultado: 13-9-2022)

Cummings, M. L. (2006). Automation and Accountability in Decision Support System Interface Design. The Journal of Technology Studies, 32(1). [PDF] Disponible en: https://scholar.lib.vt.edu/ejournals/JOTS/v32/v32n1/pdf/cummings.pdf (Consultado: 13-9-2022)

De-Arteaga, M., Fogliato, R. y Chouldechova, A. (2020). A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–12. [PDF] Disponible en: https://arxiv.org/pdf/2002.08035.pdf (Consultado: 13-9-2022)

Deloitte y Lisbon Council. (2020). Study on public sector data strategies, policies, and governance. Comisión Europea. [PDF] Disponible en: https://joinup.ec.europa.eu/sites/default/files/custom-page/attachment/2020-06/DIGIT%20-%20D01%20-%20Study%20on%20public%20sector%20data%20strategies%2C%20policies%20and%20governance%20v3annexes.pdf (Consultado: 13-9-2022)

Dietvorst, B. J., Simmons, J. P. y Massey, C. (2014). Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. SSRN Electronic Journal, 143(6), pp. 1–13. [PDF] Disponible en: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1392&context=fnce_papers (Consultado: 13-9-2022)

Digital Future Society. (2020). Diseños sensibles a las cuestiones de género en el estado de bienestar digital. [online] Disponible en: https://digitalfuturesociety.com/es/report/exploring-gender-responsive-designs-in-digital-welfare/ (Consultado: 13-9-2022)

Digital Future Society. (2021). Gobernanza y algoritmos: Riesgos y potencial del uso de la inteligencia artificial en el sector público. [online] Disponible en: https://digitalfuturesociety.com/es/report/governing-algorithms/ (Consultado: 13-9-2022)

Dourish, P. (2001). Where the Action Is: The Foundations of Embodied Interaction. [online] Disponible en: https://direct.mit.edu/books/book/3875/Where-the-Action-IsThe-Foundations-of-Embodied (Consultado: 13-9-2022).

Eiriksson Arent, B. (2019). Analyse : Udbetaling Danmarks Systematiske. Justitia og forfatteren. [PDF] Disponible en: http://justitia-int.org/wp-content/uploads/2019/07/Analyse-Udbetaling-Danmark-systematiske-overva%CC%8Agning.pdf (Consultado: 13-9-2022)

European Border and Coast Guard Agency. (2021). Artificial Intelligence-Based Capabilities for the European Border and Coast Guard. [PDF] Disponible en: https://frontex.europa.eu/assets/Publications/Research/Frontex_AI_Research_Study_2020_final_report.pdf (Consultado: 13-9-2022)

Fergusson, J. (2014). Twelve Seconds to Decide. In search of excellence: Frontex and the principle of best practice. [PDF] Disponible en: https://frontex.europa.eu/assets/Publications/General/12_seconds_to_decide.pdf (Consultado: 13-9-2022)

Goodman, B. W. (2016). Economic Models of (Algorithmic) Discrimination. 29th Conference on Neural Information Processing Systems [Prepublicación], (Nips). [PDF] Disponible en: http://www.mlandthelaw.org/papers/goodman2.pdf (Consultado: 13-9-2022)

Green, B. (2020). The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 594–606. [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3351095.3372869 (Consultado: 13-9-2022)

Green, B. (2021). The Flaws of Policies Requiring Human Oversight of Government Algorithms. SSRN Electronic Journal, pp. 1–42. [PDF] Disponible en: https://arxiv.org/ftp/arxiv/papers/2109/2109.05067.pdf (Consultado: 13-9-2022)

Green, B. y Chen, Y. (2019a). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments, pp. 90–99. [online] Disponible en: https://doi.org/10.1145/3287560.3287563 (Consultado: 13-9-2022)

Green, B. y Chen, Y. (2019b). The Principles and Limits of Algorithm-in-the-Loop Decision Making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW). [PDF] Disponible en: https://dl.acm.org/doi/pdf/10.1145/3359152 (Consultado: 13-9-2022)

Green, B. y Chen, Y. (2020). Algorithmic Risk Assessments Can Alter Human Decision-Making Processes in High-Stakes Government Contexts. [PDF] Disponible en: https://arxiv.org/pdf/2012.05370.pdf (Consultado: 13-9-2022)

Heilbrun, K., Dvoskin, J., Hart, S. y McNiel, D. (1999). Violence risk communication: Implications for research, policy, and practice. Health, Risk and Society, 1(1), pp. 91–105. [online] Disponible en: https://www.tandfonline.com/doi/abs/10.1080/13698579908407009?journalCode=chrs20 (Consultado: 13-9-2022)

Jung, H., Stolterman, E., Ryan, W., Thompson T. y Siegel, M. (2008). Toward a framework for ecologies of artifacts: how are digital artifacts interconnected within a personal life? Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges, pp. 201–210. [online] Disponible en: https://dl.acm.org/doi/10.1145/1463160.1463182 (Consultado: 13-9-2022)

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux. [online] Disponible en: https://books.google.es/books?id=ZuKTvERuPG8C (Consultado: 13-9-2022)

Kayser-Bril, N. (2020). In a quest to optimize welfare management, Denmark built a surveillance behemoth. [online] Disponible en: https://automatingsociety.algorithmwatch.org/report2020/denmark/denmark-story/ (Consultado: 22-9-2022)

Kemper, J. y Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. Information Communication and Society, 22(14), pp. 2081–2096. [online] Disponible en: https://www.researchgate.net/publication/325827444_Transparent_to_whom_No_algorithmic_accountability_without_a_critical_audience (Consultado: 13-9-2022)

Kulju, M., Ylikauppila, M., Toivonen, S. y Salmela, L. (2019). A Framework for Understanding Human Factors Issues in Border Control Automation. IFIP Advances in Information and Communication Technology. Springer International Publishing. [online] Disponible en: https://hal.archives-ouvertes.fr/hal-02264619/ (Consultado: 13-9-2022)

Lee, J. D. y See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society, 46(1), pp. 50–80. [PDF] Disponible en: https://user.engineering.uiowa.edu/~csl/publications/pdf/leesee04.pdf (Consultado: 13-9-2022)

McCallum, K. E., Boccaccini, M. T. y Bryson, C. N. (2017). The Influence of Risk Assessment Instrument Scores on Evaluators' Risk Opinions and Sexual Offender Containment Recommendations. Criminal Justice and Behavior, 44(9), pp. 1213–1235. [online] Disponible en: https://journals.sagepub.com/doi/abs/10.1177/0093854817707232 (Consultado: 13-9-2022)

Miron, M. (2018). Interpretability in AI and its relation to fairness, transparency, reliability and trust. [online] Disponible en: https://ec.europa.eu/jrc/communities/en/community/humaint/article/interpretability-ai-and-its-relation-fairness-transparency-reliability-and (Consultado: 13-9-2022)

Misuraca, G. y Noordt, C. van (2020). Report: AI Watch – Artificial Intelligence in public services | Overview of the use and impact of AI in public services in the EU. [PDF] Disponible en: https://joinup.ec.europa.eu/collection/elise-european-location-interoperability-solutions-e-government/document/report-ai-watch-artificial-intelligence-public-services-overview-use-and-impact-ai-public-services (Consultado: 13-9-2022)

Myers-West, S., Whittaker, M. y Crawford, K. (2019). Discriminating Systems. Gender, Race, and Power in AI. [PDF] Disponible en: https://ainowinstitute.org/discriminatingsystems.pdf (Consultado: 13-9-2022)

Noori, S. (2021). Suspicious Infrastructures: Automating Border Control and the Multiplication of Mistrust through Biometric E-Gates. Geopolitics, 00(00), pp. 1–23. [PDF] Disponible en: https://www.tandfonline.com/doi/pdf/10.1080/14650045.2021.1952183?needAccess=true (Consultado: 13-9-2022)

Østergaard Madsen, C., Lindgren, I. y Melin, U. (2022). The accidental caseworker - How digital self-service influences citizens' administrative burden. Government Information Quaterly. [PDF] Disponible en: https://www.sciencedirect.com/science/article/pii/S0740624X21000897 (Consultado: 13-9-2022)

Parlamento Europeo y Consejo de la Unión Europea. (2016). Reglamento general de protección de datos de la UE. [online] Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679 (Consultado: 13-9-2022)

Planas Bou, C. (2021). Catalunya usa un algoritmo para ayudar a decidir a qué presos concede la libertad condicional. El Periódico. [online] Disponible en: https://www.elperiodico.com/es/sociedad/20211117/catalunya-algoritmo-decidir-presos-concede-12859785 (Consultado: 13-9-2022)

Redden, J., Dencik, L. y Warne, H. (2020). Datafied child welfare services: unpacking politics, economics and power. Policy Studies, 41(5), pp. 507–526. [PDF] Disponible en: https://www.tandfonline.com/doi/pdf/10.1080/01442872.2020.1724928?needAccess=true (Consultado: 13-9-2022)

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 1(5), pp. 206–215. [PDF] Disponible en: https://arxiv.org/pdf/1811.10154.pdf (Consultado: 13-9-2022)

Saura, G. y Aragó, L. (2021). Un algoritmo impreciso condiciona la libertad de los presos. La Vanguardia. [online] Disponible en: https://www.lavanguardia.com/vida/20211206/7888727/algoritmo-sirve-denegar-permisos-presos-pese-fallos.html (Consultado 22-9-2022)

Scurich, N., Monahan, J. y John, R. S. (2012). Innumeracy and Unpacking: Bridging the Nomothetic/Idiographic Divide in Violence Risk Assessment. Law and Human Behavior, 36(6), pp. 548–554. [PDF] Disponible en: https://www.researchgate.net/publication/224869265_Innumeracy_and_Unpacking_Bridging_the_NomotheticIdiographic_Divide_in_Violence_Risk_Assessment (Consultado: 13-9-2022)

Seaver, N. (2019). Knowing Algorithms. digitalSTS, pp. 412–422. [PDF] Disponible en: https://digitalsts.net/wp-content/uploads/2019/03/26_Knowing-Algorithms.pdf (Consultado: 13-9-2022)

Stevenson, M. y Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. SSRN Electronic Journal [Prepublicación]. [PDF] Disponible en: https://docs.iza.org/dp12853.pdf (Consultado: 13-9-2022)

Suchman, L. A. (2007). Human-Machine Reconfigurations: Plans and Situated Actions. 2nd Edition. Nueva York, EE. UU.: Cambridge University Press.

Tan, S., Adebayo, J., Inkpen, K. y Kamar, E. (2018). Investigating Human + Machine Complementarity for Recidivism Predictions. [PDF] Disponible en: https://arxiv.org/pdf/1808.09123.pdf (Consultado: 13-9-2022)

Van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M. y Kostakos, V. (2019). Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW). [PDF] Disponible en: https://dl.acm.org/doi/abs/10.1145/3359130 (Consultado: 13-9-2022)

Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy and Internet, 11(1), pp. 104–122. [PDF] Disponible en: https://onlinelibrary.wiley.com/doi/epdf/10.1002/poi3.198 (Consultado: 13-9-2022)

Zhang, Y., Vera Liao, Q. y Bellamy, K. E. (2020). Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. FAT* 2020 – Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305. [PDF] Disponible en: https://arxiv.org/pdf/2001.02114.pdf (Consultado: 13-9-2022)

Zuiderwijk, A., Chen, Y. C. y Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly (May 2020), p. 101577. [PDF] Disponible en: https://www.researchgate.net/publication/350333329_Implications_of_the_use_of_artificial_intelligence_in_public_governance_A_systematic_literature_review_and_a_research_agenda (Consultado: 13-9-2022)

Agradecimentos

Autor principal

Manuel Portela es investigador posdoctoral en el Grupo de Investigación Ciencia Web y Computación Social (Universidad Pompeu Fabra). Se ha especializado en diseño de interacciones y geografía humana. Sus investigaciones se centran en la justicia algorítmica y en cómo afecta la interacción entre personas y máquinas al uso de la IA en la toma de decisiones. Antes de iniciar su trayectoria académica, dirigió varios proyectos de transformación digital en la Administración local de Buenos Aires (Argentina).

Coautora

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.

Equipo de Digital Future Society Think Tank

Gracias a las siguientes compañeras de Digital Future Society Think Tank por sus aportaciones y su apoyo en la elaboración de este informe:

- **Carina Lopes**, directora de Digital Future Society Think Tank
- **Olivia Blanchard**, investigadora de Digital Future Society Think Tank

Citas

Este informe se debe citar de la siguiente manera:

- Digital Future Society. (2022). Hacia una supervisión significativa de los sistemas automatizados de toma de decisiones. Barcelona, España.

Datos de contacto

Si desea ponerse en contacto con el equipo de Digital Future Society Think Tank, envíe un correo electrónico a thinktank@digitalfuturesociety.com


Digital
Future Society