Revista de Estudios Ius Novum
Vol. XVI N° 1 año 2023
[pp. 221 - 274]

TRANSPARENCIA EN LOS PROCESOS DE ELABORACIÓN AUTOMATIZADA DE PERFILES DE RIESGO DE FRAUDE: ANÁLISIS DE LA REGULACIÓN EUROPEA A LA LUZ DEL CASO SYRI EN LOS PAÍSES BAJOS

[Transparency in automated processes for developing fraud risk profiles: Analysis of European regulation in light of the SyRI case in the Netherlands]

DIEGO OJEDA CIFUENTES[1]

RESUMEN

Este trabajo delimita la aplicabilidad del principio de transparencia en la elaboración automatizada de perfiles de riesgo de fraude a través de un caso de estudio: "SyRI". El análisis revela una tensión con la opacidad estratégica necesaria para evitar el juego con el sistema. Se sugiere una posición intermedia que proporcione una explicación atenuada a los titulares de datos para salvaguardar la funcionalidad del sistema, restringiendo la transparencia total a las evaluaciones de impacto y auditorías periódicas.

PALABRAS CLAVE

RGPD – Transparencia Algorítmica – Elaboración Automatizada de Perfiles – Aprendizaje Automático

ABSTRACT

This paper illustrates the applicability of the transparency principle in automated fraud risk profiling by means of a case study: "SyRI". The analysis reveals a tension with the strategic opacity necessary to avoid gaming the system. A middle position is proposed, which provides mitigated explanation to data subjects to ensure the functionality of the system, while limiting full transparency to impact assessments and periodic audits.

KEYWORDS

GDPR – Algorithmic Transparency – Profiling – Machine Learning.

[1] Estudiante de Derecho en Pontificia Universidad Católica de Valparaíso. Ayudante en Programa de Derecho, Inteligencia Artificial y Tecnología PUCV. Correo: dojedacifuentes@gmail.com

INTRODUCCIÓN

Con el auge de las técnicas de elaboración automatizada de perfiles basada en aprendizaje automático (en adelante "ML", por sus siglas en inglés), el principio de transparencia es cada vez más relevante. Nunca —en la historia de las sociedades humanas— ha existido una combinación tan compleja y ubicua de técnicas de perfilamiento e identificación como en la era del Big Data.

El contexto sociotécnico del Big Data, caracterizado principalmente por la inconmensurable disponibilidad de datos, no sólo es "grande" por su tamaño, sino también en virtud de los avances tecnológicos que representan una oportunidad clave para explotar la fuerza productiva de los datos y producir nuevo conocimiento. Pero la metamorfosis de las técnicas de vigilancia hacia procesos basados en ML también trae consigo un cambio cualitativo en su propósito, desafiando la noción clásica de transparencia y abriendo paso, a su vez, a nuevas formas de afectación al derecho a la protección de datos personales.

En el ámbito de la seguridad social —aunque no exclusivamente—, resulta especialmente preocupante cómo los Estados con más recursos están usando cada vez más este tipo de tecnologías para "automatizar, predecir, identificar, vigilar, detectar, singularizar y castigar"[2]. Ciertamente, las posibilidades de optimizar los recursos fiscales son enormes. Pero pese a ser una empresa legítima, las asimetrías de información y la opacidad —que suelen ser la regla general en esta materia— se proyectan como una verdadera "caja negra" para la mayoría de los ciudadanos, amenazando las posibilidades de repercutir positivamente al obliterar correlaciones basadas en información sensible, sesgos, discriminación, intrusiones ilegítimas a la privacidad, etc.

Resulta manido a estas alturas insistir en que una de las tareas más urgentes y prioritarias para una implementación exitosa de estos sistemas es, ineludiblemente, la transparencia algorítmica. Como han tenido ocasión de señalar Lilian EDWARDS y Michael VEALE, existe una creciente preocupación en la literatura sobre avanzar hacia un fútil paradigma de transparencia sin sentido[3], que pueda ser percibido como esencialmente benigno, cuando en realidad se estuviera tejiendo un Velo de Maya que obture, y de paso legitime, sistemas altamente intrusivos y potencialmente discriminatorios. Desafortunadamente, esta cavilación no parece ser del todo descaminada; cuando de ML se trata, la transparencia es siempre un dédalo multidimensional, en perpetuo movimiento y en el cual interactúan una plétora factores contextuales difíciles de canalizar.

La discusión ha adquirido un renovado interés en la literatura a partir del reciente caso NJCM et al contra los Países Bajos[4], en que el Tribunal de Distrito de la Haya declaró que la normativa y despliegue de un algoritmo empleado para la elaboración automatizada de perfiles de riesgo de fraude a la seguridad social –el sistema de indicación de riesgo neerlandés "SyRI" (*Systeem Risico Indicatie*)– sería contrario al artículo 8 del Convenio Europeo para la Protección de los Derechos

[2] Organización de Naciones Unidas, "*Informe del relator especial sobre la pobreza extrema y los derechos humanos"* 74 período de sesiones, punto 72 (b) del orden del día provisional, A/74/48.037 del 11 de octubre de 2019, p. 4. [Visible en: https://undocs.org/pdf?symbol=es/A/74/493] [Consultado por última vez: 9/08/2021].

[3] EDWARDS, Lilian - VEALE, Michael, *Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for*, en *Duke L. & Tech. Rev.* 16 (2017) 18, pp. 81-82.

[4] Sentencia del 5 de febrero de 2020 del Tribunal de Distrito de la Haya (Rechtbank Den Haag) N° C/09/550982/ HA ZA 18/388. [Visible en https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:865] [Consultado por última vez: 9/08/2021].

Humanos y de las Libertades Fundamentales (en adelante, CEDH), basando buena parte del peso de su conclusión en la aplicación del principio de transparencia. Dicho fallo, calificado como un "precedente histórico" por la ONU[5], se trata nada menos que del primer precedente jurisprudencial en pronunciarse contrario al uso gubernamental de un algoritmo de tales características, estableciendo a su paso un alto estándar de garantías en materia de transparencia. No obstante, hasta la fecha ha habido muy poco consenso sobre cómo el principio de transparencia debe desarrollarse en aplicaciones concretas.

El quid del asunto, al menos en lo que se refiere a la vigilancia y la lucha contra el fraude, radica en que cierto grado de opacidad será habitualmente deseable para evitar el juego estratégico con el sistema (*gaming the algorithm*), de modo tal, que, resulta inmensamente controvertido determinar cómo conciliar este fin con las obligaciones en materia de transparencia. Así, por ejemplo, cuando pedimos ver en (o a través de) la "caja negra", cabe preguntarse: ¿De qué tipo de transparencia se está hablando? ¿Cuáles son sus limitaciones? ¿Acaso estaremos pidiendo demasiado? ¿Hasta qué punto deberíamos estar dispuestos a sacrificar algo de eficiencia en aras de la transparencia?

Una coyuntura adicional pasa porque el ritmo del desarrollo tecnológico supera la velocidad con que se implementan mecanismos de gobernanza algorítmica. Esto se aprecia con meridiana claridad en el contexto nacional, donde la ley que regula la protección de datos personales data del año 1999[6]. En esta línea, sin embargo, también es posible avizorar algunas iniciativas dignas de mención. El pasado 12 de diciembre de 2020 se abrió a consulta pública por el Ministerio de Ciencia, Tecnología, Conocimiento e Innovación el primer Borrador de Política Nacional de IA con recomendaciones en materia de inteligencia artificial, en cuyo contenido se recoge una propuesta de principios para la aplicación, uso y despliegue de la IA, entre los cuales se menciona reiteradamente la "transparencia y explicabilidad"[7]. En el ámbito internacional —se debe mencionar también— la Organización para la Cooperación y Desarrollo Económico (OCDE) suscribió, a finales de mayo de 2019, los Principios sobre la IA, entre los cuales también se incluye la transparencia algorítmica[8]. Se trata, pues, de una zona gris de nuestra legislación, que ante la ausencia de una matriz conceptual *sub specie saeculi*, se encuentra en un momento crítico para adecuar su marco regulatorio a estándares internacionales.

Este no es precisamente el caso de la Unión Europea. Como no podía ser de otra forma, el Reglamento General de Protección de Datos (2016) de la UE (en adelante, RGPD)[9] —comúnmente reconocido como el máximo estándar de regulación al que se puede aspirar en materia de protección de datos personales— aborda explícitamente el efecto sobre los "*derechos y libertades fundamentales de las personas físicas*" (párrafo 2 del artículo 1) que puede tener la "*elaboración*

[5] ONU, *"Landmark Dutch court ruling halts government attempts to spy on the poor: UN expert*", Comunicado de prensa, 5 de febrero de 2020. [Visible en: https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25522] [Consultado por última vez: 9/08/2021].

[6] Ley N° 19.628 Sobre Protección de la Vida Privada.

[7] Vid. Consulta Pública, *"Política nacional de inteligencia artificial*" (2020). [Visible en: https://www.minciencia.gob.cl/legacy-files/borrador_politica_nacional_de_ia.pdf] [Consultado por última vez: 9/08/2021].

[8] Vid. OCDE, "*Recomendación del Consejo de Inteligencia Artificial*" (2019). [Visible en: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449] [Consultado por última vez: 9/08/2021].

[9] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo de 27 de abril de 2016 relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

automatizada de perfiles" (artículo 4, apartado 4), incluyendo en un extenso articulado todo un sistema de transparencia calificado. De este modo, el RGPD —al ser el primer instrumento legislativo que regula explícitamente la elaboración automatizada de perfiles— sienta un importante precedente en la materia: su éxito (o fracaso) puede tener repercusiones que se extiendan mucho más allá de Europa.

A fortiori, dado el embrionario estado de desarrollo en que se encuentra esta materia, resulta fundamental analizar la eficacia del marco regulatorio europeo, y especialmente el RGPD, para mitigar los efectos de la opacidad algorítmica y promover algoritmos más responsables. Dado el reciente precedente jurisprudencial generado a raíz de SyRI, un punto de partida interesante para abordar esta problemática es la seguridad social. Y es que, si bien los desafíos legales que trae consigo la elaboración de perfiles algorítmicos varían según dónde y con qué propósito se utilicen, las consecuencias del desarrollo tecnológico pueden observarse a nivel global, de modo que, al dilucidar hasta qué punto el marco europeo establece mecanismos de protección efectivos frente a la opacidad, el estudio contribuye a la búsqueda de una regulación que permita articular la protección de las libertades y derechos de las personas físicas frente a la automatización, en lo que se refiere a tecnologías que pueden resultar opacas, excesivamente intrusivas y potencialmente discriminatorias.

HIPÓTESIS Y OBJETIVOS

En el contexto del marco de protección de datos de la UE, esta memoria postula que la opacidad algorítmica, en la elaboración automatizada de perfiles de riesgo de fraude a la seguridad social, puede vulnerar el derecho a la protección de datos personales, especialmente si faltan garantías compensatorias suficientes.

Así, el objetivo principal de este trabajo es delimitar el despliegue de la fuerza normativa del principio de transparencia algorítmica aplicado a la elaboración automatizada de perfiles de riesgo para la detección de fraude en la seguridad social. Para la concreción de este objetivo, se pretende:

a) Explorar qué es la elaboración automatizada de perfiles y cómo se aplica a la lucha contra el fraude en la seguridad social utilizando técnicas de ML.

b) Identificar los componentes esenciales de la transparencia algorítmica en el proceso de perfilado y reconocer los tipos de opacidad algorítmica, explicando sus causas y efectos.

c) Examinar las soluciones posibles para enfrentar la opacidad algorítmica en la regulación de protección de datos de la UE.

d) Estudiar la aplicabilidad práctica del principio de transparencia en el caso SyRI, señalando sus limitaciones y potenciales soluciones legales y técnicas.

METODOLOGÍA Y RUTA DE INVESTIGACIÓN

El presente estudio se fundamenta en una amalgama de metodologías complementarias: el método sistémico-estructural-funcional, el estudio de caso y un enfoque interdisciplinario. A través del método sistémico-estructural-funcional, se desglosa el objeto de investigación, permitiendo una comprensión integral del fenómeno de la transparencia como un elemento relacional. Esto implica un análisis que contempla diversos factores contextuales, como los

distintos tipos de receptores y las diversas etapas del proceso de elaboración automatizada de perfiles, para así determinar las implicancias específicas del fenómeno de la opacidad algorítmica. El sistema antifraude neerlandés SyRI se utiliza como estudio de caso, ofreciendo un escenario práctico para contextualizar las disposiciones legales. Adicionalmente, se aplica una perspectiva interdisciplinaria que aborda aspectos técnicos del Machine Learning, contribuyendo a la comprensión de la realidad tecnológica subyacente. En conjunto, estas metodologías permiten un análisis riguroso, detallado y enriquecedor de la problemática estudiada.

El desarrollo de este trabajo se estructura en 5 secciones. Tras las observaciones introductorias, comenzamos esta tarea abordando las generalidades de la elaboración automatizada de perfiles; para ello, teniendo en cuenta las diversas conceptualizaciones propuestas, se analiza brevemente el proceso de perfilado bajo la óptica del ML y se da cuenta de su aplicación concreta al ámbito de la detección del fraude a la seguridad social a través de SyRI. El segundo apartado analiza, a partir de la literatura especializada, por un lado, el principio de transparencia y sus fundamentos, y, por otro, el fenómeno de la opacidad, sus causas y consecuencias. Estas dos primeras secciones dan cuenta de los presupuestos teóricos necesarios para la articulación del debate jurídico que sigue. A partir de las categorías analizadas, el apartado III desarrolla el encaje jurídico y las implicancias normativas de la elaboración automatizada de perfiles y la transparencia en la legislación de la UE sobre protección de datos, al mismo tiempo que conecta los dos temas. Dicho análisis se circunscribe principalmente a la regulación establecida por el RGPD. El cuarto epígrafe representa el momento de síntesis de esta memoria, después del trabajo descriptivo y analítico realizado en las partes anteriores. De este modo, se desarrolla el problema de las posibilidades de aplicación concreta del principio de transparencia algorítmica frente al tipo de opacidad estratégica. Este problema se aborda, en primer lugar, desde la óptica del operador jurisdiccional en el caso NJCM et al contra los Países Bajos y, en segundo término, evaluando de manera autónoma la aplicabilidad del principio de transparencia, identificando vías positivas para conseguir este objetivo y las limitaciones de este enfoque. Y para finalizar, como es la tradición, algunas conclusiones.

I. ASPECTOS GENERALES DE LA ELABORACIÓN AUTOMATIZADA DE PERFILES

Esta sección tiene como propósito, proporcionar una adecuada comprensión del objeto de estudio, sus partes internas y su concreto ámbito de aplicación a la detección del fraude a la seguridad social. Así, se abordarán tres puntos: en primer lugar, la conceptualización de la elaboración automatizada de perfiles de riesgo; en segundo lugar, una breve descripción sobre cómo se aplican las técnicas de ML al proceso de perfilado; y para concluir esta sección, se presenta el caso de estudio, *i.e.* la herramienta SyRI. Ello nos permitirá tener claridad respecto de los presupuestos conceptuales sobre los cuales articular el examen sobre la opacidad y transparencia, cuya elucidación se somete a examen en la sección siguiente.

1. ¿Qué es la elaboración automatizada de perfiles riesgo?

El artefacto del Estado moderno, desde sus orígenes, ha reclamado una insaciable necesidad de perfilar a los alineados a su territorio[10]. Y desde Jeremy BENTHAM, sabemos que las prácticas de vigilancia estatal operan creando un estado de visibilidad constante en el sujeto observado, permitiendo así la "automatización del poder"[11]. En esta línea, la elaboración automatizada de perfiles suele describirse como una práctica panóptica gubernamental específica; "la vigilancia de datos"[12]. Se trata, pues, de un tipo de vigilancia que se caracteriza "por el monitoreo, recopilación y procesamiento de datos para el control continuo de los ciudadanos"[13]. En consecuencia, la elaboración automatizada de perfiles es, ante todo, una forma de minería de datos[14].

Desde un punto de vista jurídico, la noción de "elaboración de perfiles" durante mucho tiempo estuvo vetada de un estatus legal claramente definido en el derecho europeo. De hecho, pese a ser una práctica de larga data, hasta antes de la entrada en vigor del RGPD no existió una regulación que pudiese conceptualizar esta figura. No obstante, dicha noción era reconducible a la regulación genérica establecida por la antigua Directiva 95/46/CE sobre protección de datos (en adelante, DPD), que definía en su artículo 2, lit. b) "*tratamiento de datos*" como: "*cualquier operación o conjunto de operaciones efectuadas o no mediante procedimientos automatizados y aplicadas a datos personales, como la recogida, registro, organización, conservación, elaboración o modificación, extracción, consulta, utilización, comunicación por transmisión, difusión o cualquier otra forma que facilite el acceso a los mismos, cotejo o interconexión; así como su bloqueo, supresión o destrucción*" [15].

Así, una de las novedades que trajo consigo el RGPD es la introducción de una definición legal de "elaboración de perfiles". De este modo, en la terminología empleada por este instrumento, la elaboración de perfiles es: "*toda forma de tratamiento automatizado de datos personales consistente en utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular para analizar o predecir aspectos relativos al rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física*"[16]. Asimismo, algunos tipos de perfiles se ejemplifican en el Considerando 71, según se refieran a: "*aspectos relacionados con el rendimiento en el trabajo, la situación económica, la salud, las preferencias o intereses personales, la fiabilidad o el comportamiento, la situación o los movimientos del interesado*".

---

[10] HILDEBRANDT, Mireille, *The dawn of a critical transparency right for the profiling era*, en BUS, Jacques - HILDEBRANDT, Mireille (editores), *Digital enlightenment yearbook* (Amsterdam, 2012), p. 42.

[11] FOUCAULT, Michael, *Vigilar y castigar: nacimiento de la prisión* (Traducción de Aurelio Garzón del Camino, Siglo XXI, Editores, Buenos Aires, 2002), p. 185 en GARRIGA, Ana, *La elaboración de perfiles y su impacto en los derechos fundamentales: una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea*. en *Derechos y Libertades* 38 (2018) 2, p. 122.

[12] VAN DIJCK, José, *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology,* en *Surveillance & society* 12 (2014) 2, p. 210, en BÜCHI, Moritz - FOSCH-VILLARONGA, Edward - LUTZ, Christoph - TAMO-LARRIEUX, Aurelia - VELIDI, Shruthi - VILJORN, Salomé, *Chilling effects of profiling activities: Mapping the issues,* en *Computer Law & Security Review* 36 (2020), p. 3.

[13] BÜCHI, Moritz - FOSCH-VILLARONGA, Edward - LUTZ, Christoph - TAMO-LARRIEUX, Aurelia - VELIDI, Shruthi - VILJORN, Salomé, *Chilling effects of profiling activities: Mapping the issues,* en *Computer Law & Security Review* 36 (2020), p. 3.

[14] HILDEBRANDT, M., *Defining profiling: a new type of knowledge?,* en HILDEBRANDT, Mireille - GUTWIRTH, Serge (editores), *Profiling the European citizen* (Springer, Dordrecht, 2008), pp. 17-45.

[15] Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. Artículo 2, lit. b).

[16] Artículo 4, apartado 4 del RGPD.

Existe una vasta lista de conceptos relativos a la elaboración de perfiles en la literatura especializada. Entre ellas, pese a que no existe un concepto fijo e indiscutible, la definición que probablemente goce de más amplia aceptación se encuentra en la obra de HILDEBRANDT, quien entiende esta actividad como "el proceso de descubrir correlaciones entre datos en bases de datos que se pueden usar para identificar y representar a un sujeto humano o no humano (individual o grupal) y/o la aplicación de perfiles (conjuntos de datos correlacionados) para individualizar y representar a un sujeto o para identificar a un sujeto como miembro de un grupo o categoría"[17]. Como es posible advertir, se trata de una definición más amplia que la anterior.

Así, en contraste con la definición de HILDEBRANDT, podemos señalar que la definición del RGPD consiste en una forma específica de elaboración de perfiles, puesto que: i) se trata de una forma automatizada de elaborar perfiles; ii) implica el uso de datos personales; y iii) tiene como objetivo personas físicas. A fin de delimitar correctamente nuestro objeto de estudio (*i.e.* la elaboración automatizada de perfiles de riesgo de fraude), vale la pena establecer estas precisiones conceptuales con algún grado de detalle.

Como primera aproximación, es preciso subrayar que los métodos actuariales empleados para descubrir patrones pueden clasificarse, según el grado de participación humana, en 3 grandes grupos: no automatizados, semiautomatizados y autónomos[18]. En esta idea, los modelos de los cuales se ocupan las siguientes líneas son los referentes al campo del ML. En consecuencia, para efectos de este estudio, se excluyen aquellos que no implican algún tipo de razonamiento automatizado. En lo que se refiere a los otros dos, la diferencia suele ser mucho más sutil y controversial, pues depende de la valoración del grado de participación humana dentro del bucle algorítmico. Volveremos sobre este punto al analizar el artículo 22 del RGPD.

Asimismo, tal como se desprende de la definición anterior, la elaboración de perfiles no necesariamente implica inferencias sobre personas físicas (*e.g.* lugares, eventos u objetos de interés). En el ámbito de los perfiles automatizados de riesgo, podemos tomar como ejemplo el que nos presenta VAN SCHENDEL: el Sistema de Anticipación del Crimen (CAS), usado por la policía de los Países Bajos con el objeto de determinar el nivel de riesgo de un área específica, sin identificar a una persona determinada; el sistema crea una cuadrícula, actualizada cada 14 días, indicando qué delito es más probable que ocurra a una determinada hora, según el área objetivo[19]. Ahora bien, sin perjuicio de que esta modalidad puede traer como consecuencia que, indirectamente, la inferencia de riesgo sobre un área se vincule a una persona o grupo de personas físicas, en lo que sigue sólo se hará referencia a dicha actividad cuando ésta tiene por objeto una inferencia vinculada directamente a personas físicas.

En cuanto a su finalidad, como ya se ha planteado, se consideran aquí las técnicas de elaboración de perfiles que se utilizan para identificar categorías o grupos de personas con una finalidad concreta: evaluar el nivel de riesgo de cometer fraude a la seguridad social. Se trata, sin embargo, de inferencias probabilísticas. El significado de ese riesgo asociado a una persona física,

---

[17] HILDEBRANDT, Mireille, *Profiling and AML,* en RANNENBERG, Kai - ROYER, Denis – DEUKER, André (editors), *The Future of Identity in the Information Society. Challenges and Opportunities* (Berlín, Springer-Verlag Berlin Heidelberg, 2009), p. 275.

[18] BOSCO, Francesca - CREEMERS, Niklas - FERRARIS, Valeria - GUAGNIN, Daniel - KOOPS, Bert-Jaap, *Profiling technologies and fundamental rights and values: regulatory challenges and perspectives from European Data Protection Authorities* en GUTWIRTH, Serge - LEENES, Ronald - DE HERT, Paul (editores), *Reforming European data protection law* (Dordrecht, Springer-Verlag Berlin Heidelberg, 2015), p. 8.

[19] VAN SCHENDEL, Sascha, *The challenges of risk profiling used by law enforcement: Examining the cases of COMPAS and SyRI*, en REINS, Leonie (editor), *Regulating New Technologies in Uncertain Times* (The Hague, 2019), p. 230.

en general, se traducirá como la probabilidad de que en el individuo se produzca cierta conducta, la cual tendrá, según el ámbito de aplicación, un significado diferente. Esa actividad produce, pues, un tipo de "conocimiento inductivo"; pese a que las correlaciones sólo contienen información sobre si el patrón de desviación de una media es similar para dos variables de interés, sin referencia a una causa concreta, aún se puede estimar "la probabilidad de que las cosas salgan igual en el futuro"[20].

Una última distinción relevante es la que plantea VAN SCHENDEL a propósito del modo en que se despliega el análisis del riesgo respecto de los individuos: "individual" y "general"[21]. La elaboración de perfiles de riesgo se aplica de manera individual cuando ya existe un individuo determinado a quien se dirige el análisis de riesgo[22]. Podemos citar como ejemplo de ello al funcionamiento del sistema HART, empleado por la policía de Durham, Reino Unido, para predecir la probabilidad de reincidencia de los reclusos[23]. En tal caso, se aplica un perfil de riesgo a un sujeto determinado, quien anteriormente ha cometido un delito, con el fin de evaluar la probabilidad de que reincida en el futuro. Se trata, por ende, de un tipo de vigilancia reactiva. En el segundo caso, en cambio, los algoritmos se destinan a encontrar personas o grupos que sean interesantes de investigar, pues como indica VAN SCHENDEL, este uso de la elaboración de perfiles "conduce al punto de partida de una investigación"[24]. Es decir, en lugar de buscar información sobre una persona específica, el sistema crea una categorización que especifica qué individuos son de un determinado riesgo, sobre la base de un modelo de riesgo predeterminado[25]. Así, en contraste del caso anterior, se trata de un tipo de vigilancia proactiva. Este segundo tipo de despliegue es el que se abordará a lo largo de esta memoria, ilustrado a través del mencionado sistema antifraude SyRI.

2. El proceso de perfilado bajo la óptica del ML

El ML es un subcampo de la inteligencia artificial (IA) que se ocupa del diseño de técnicas estadísticas para la clasificación de patrones mediante el uso de algoritmos de aprendizaje. Por consiguiente, su uso es perfectamente compatible con la elaboración de perfiles de riesgo. En concreto, su función es decidir a cuál categoría, de un conjunto más amplio de categorías (definidas por las unidades de salida, e.g. alto riesgo o bajo riesgo), pertenece una entrada determinada[26]. Nótese que la idea de aprendizaje no alude en modo alguno a algún tipo de representación conceptual, sino al "aumento del rendimiento del software en una tarea específica" a partir de ejemplos en bases de datos[27]. Se trata, por ende, de programas informáticos que se optimizan (o aprenden) en tareas de clasificación, a partir de patrones observados en bases de datos con ejemplos anteriores[28].

---

[20] HILDEBRANDT, cit. (n. 13), p. 18.
[21] VAN SCHENDEL, cit. (n. 18), pp. 229-231.
[22] Ibid., p. 231.
[23] Vid. OSWALD, Marion, *Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality*, en *Information & Communications Technology Law* 27 (2018) 2, pp. 223-250.
[24] VAN SCHENDEL, cit. (n. 18), p. 229.
[25] Ibid.
[26] MARCUS, Gary, *Deep learning: A critical appraisal* (2018), p. 9. [Visible en: https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf] [Consultado por última vez: 9/08/2021].
[27] EDWARDS - VEALE, cit. (n. 2), p. 26.
[28] BROUSSARD, Meredith, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2019), p. 93.

Las técnicas algorítmicas de clasificación destinadas a la elaboración de perfiles de riesgo pueden incluir una plétora de modelos diferentes, verbigracia: árboles de decisión, redes neuronales, redes bayesianas, etc. Para maximizar la precisión, los modelos también se pueden combinar en "conjuntos de modelos" (e.g. combinación de árboles de decisión mediante ensamble, creando un bosque aleatorio), un enfoque que se usa a menudo en las competencias de ML[29]. Amén de un modelo específico, y pese a los no menores matices entre unos y otros, es posible identificar en todos ellos 3 fases en común: a) recolección de datos; b) construcción del modelo; y c) implementación del modelo[30]. Siguiendo este esquema, a continuación, se ofrece una breve explicación de este proceso.

La primera fase consta de la recopilación de conjuntos de datos que deben servir como entrada para su tratamiento ulterior[31]. Usualmente, los conjuntos de datos recopilados en esta fase se dividen en un subconjunto de entrenamiento (80% de los datos) y otro para probar la precisión en la clasificación (20% de los datos)[32]. Esta fase comprende, no sólo la obtención de los datos "en bruto", sino también un costoso proceso de limpieza, etiquetado y clasificación, para ser utilizados posteriormente. De este modo, la "variable objetivo" define lo que se busca predecir o clasificar, es decir, "los resultados de interés" [33] que en nuestro caso serían los indicadores de riesgo relevantes que deben tenerse en cuenta para la clasificación de fraude. En cambio, las "etiquetas de clase dividen todos los posibles valores de la variable objetivo en categorías mutuamente excluyentes"[34].

En virtud del patrón de aprendizaje empleado, es posible clasificar los algoritmos de ML en 2 grandes categorías: aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, se utiliza un "vector de variables", y una "etiqueta correcta" para ese vector, de modo que el algoritmo buscará predecir esta "etiqueta correcta" a partir de las variables de entrada[35]. En otras palabras, el algoritmo utiliza una muestra de datos etiquetados para aprender una regla general que convierte a las entradas en salidas. Esta será la regla general en el caso de la elaboración de perfiles de riesgo. En el aprendizaje no supervisado, por otro lado, se trata de ver qué características están agrupadas, sin saber previamente qué podrían significar[36]. Es decir, el algoritmo identifica patrones ocultos a partir de datos no etiquetados.

En lo que se refiere a la construcción del modelo, podemos distinguir dos operaciones paralelas: "aprendizaje y clasificación"[37]. Los algoritmos de aprendizaje primero se entrenan con datos de prueba, dando como resultado una matriz de ponderaciones. Luego, ésta será utilizada por el clasificador para determinar la categorización de nuevos datos de entrada. Al finalizar este proceso, se ha creado un modelo en que, a partir de los pesos generados durante el entrenamiento, el algoritmo clasificador podrá ser consultado con variables de entrada, denominadas como "conjunto de características", y producir una salida, vale decir, una

[29] BURRELL, Jena, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, en *Big Data & Society* 3 (2016) 1, p. 5.
[30] DE LAAT, Paul, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?*, en *Philosophy & Technology* 31 (2018) 4, pp. 527 y ss. [Visible en: https://link.springer.com/article/10.1007/s13347-017-0293-z] [Consultado por última vez: 9/08/2021].
[31] Ibid., p. 530.
[32] Ibid., p. 531.
[33] BAROCAS, Solon - SELBST, Andrew, *Big data's disparate impact,* en *California Law Review* 104 (2016) 3, p. 678 y ss.
[34] Ibid.
[35] EDWARDS - VEALE, cit. (n. 2) pp. 25-26.
[36] Ibid.
[37] BURRELL, cit. (n. 28), p.5.

"categoría"[38]. Por ejemplo, un algoritmo para la clasificación del riesgo de fraude tomaría un conjunto de características extraídas de los datos personales de una persona o grupo (*e.g.* código postal, beneficios sociales recibidos, historial de saldo bancario, etc.), y produce una categoría de salida, *id est* una determinada puntuación de riesgo que se vincula a ese individuo o grupo (*e.g.* alto o bajo riesgo de cometer fraude).

La elección entre un modelo u otro dependerá, en cierta medida, de diversos factores "tales como el dominio, la precisión demostrada en la clasificación y los recursos computacionales disponibles"[39]. No obstante, dicha elección, para efectos de la interpretabilidad del modelo, no es inocua. Existe un amplio espectro que fluctúa entre los modelos totalmente diáfanos y los no transparentes o intrínsecamente opacos. En un primer nivel están aquellos en que entendemos cómo se relacionan conjuntamente todas las variables para producir una salida[40]. Piénsese, por ejemplo, en un árbol de decisiones, un tipo de aprendizaje supervisado que se puede explicar mediante modelos lógicos del tipo "si la condición A es verdadera o las condiciones B y C son falsas, entonces predice sí, de lo contrario predice no"[41]. En un nivel de transparencia "intermedia" se encuentran modelos más complejos, tales como el "bosque aleatorio" (*random forest*), como el que utiliza el mencionado sistema HART. Y, en tercer término, podemos mencionar las redes neuronales profundas, modelos intrínsecamente opacos donde es muy difícil o imposible producir una explicación deductiva de cómo interactúan todas las variables entre sí para producir un resultado[42]. En tales casos, como observa GOODMAN, el desafío de la interpretabilidad puede ser doble: en primer lugar, "la mera complejidad del modelo subyacente" y, en segundo lugar, el hecho de que "muchas variables de predicción no tienen ningún tipo de interpretación teórica"[43].

Por último, la fase de implementación del modelo comprende el contexto que va desde la aplicación del perfil de riesgo a un individuo determinado, hasta el momento en que la inferencia arrojada por el algoritmo es utilizada para la toma de una decisión. Esta fase puede subdividirse en 4 sub etapas: primero, la recopilación de los datos, personales o no, que se van a someter a análisis; en segundo lugar, el análisis de los datos a través de los modelos de riesgo; en tercer lugar, el momento en el que el algoritmo arroja la inferencia de riesgo; y finalmente, la toma de decisión posterior, que comprende diversas posibilidades (*e.g.* donde la toma de decisiones se realiza con nula intervención humana, sistemas en los que la inferencia es un mero apoyo al juicio humano, etc.) [44].

3. Caso de estudio: Systeem Risico Indicatie (SyRI)

Debe reconocerse que las técnicas de perfilado a través de ML pueden lograr un alto grado de eficiencia en tareas de prevención y detección del fraude, que es, precisamente, uno de los

---

[38] Ibid.
[39] Ibid.
[40] RUDIN, Cynthia, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,* en *Nature Machine Intelligence* 1 (2019) 5, pp. 9 - 10.
[41] Ibid., p. 10.
[42] BURRELL, cit. (n. 28), p. 7.
[43] GOODMAN, Bryce, *A step towards accountable algorithms?: Algorithmic discrimination and the European Union general data protection,* en *29th Conference on Neural Information Processing Systems,* Barcelona, NIPS Foundation (2016), p. 4.
[44] Vid. LAZCOZ, Guillermo, *Modelos algorítmicos, sesgos y discriminación, ponencia presentada en IX Fórum de expertos y jóvenes investigadores en Derecho y Nuevas Tecnologías* (2020) [Visible en: https://www.researchgate.net/publication/338622994_Modelos_algoritmicos_sesgos_y_discriminacion].

pilares esenciales del diseño de cualquier sistema de asistencia y seguridad social. Pero pese al desaforado —y a veces casi irreflexivo— optimismo que a veces se exhibe en nombre de la eficiencia, huelga decir que, en ocasiones, tales prácticas pueden tener efectos adversos en la posición jurídica de los ciudadanos. Para estos efectos, resulta ilustrativo el sistema antifraude neerlandés "SyRI". Como se adelantaba en la introducción, se trata de uno de los casos más conspicuos dentro de esta tendencia, en que se utiliza este tipo de herramientas para elaborar perfiles de riesgo con el fin de detectar, prevenir y combatir el fraude a la asistencia y seguridad social[45]. A continuación, se aborda un breve sumario de la técnica utilizada en el despliegue de SyRI, el contexto en que se desarrolla y sus principales características.

Las bases legales para el funcionamiento SyRI están en los artículos 64 y 65 de la Ley SUWI (ley de organización de implementación y estructura de ingresos)[46] y el capítulo 5.a del Decreto SUWI (que establece normas y procedimientos)[47]. Sin embargo, el desarrollo de esta herramienta se remonta al año 2003 —casi una década antes de que existiera una base legal para su funcionamiento—, con la fundación de los equipos de intervención del Comité Directivo Nacional, en virtud del cual se buscaba coordinar la lucha contra diversos ilícitos de fraude[48]. Asimismo, las primeras experiencias con el despliegue de técnicas automatizadas para combatir el fraude datan del 2006[49]. En 2007, en respuesta a las preocupaciones planteadas por la autoridad de protección de datos, se creó un entorno seguro denominado "método *blackbox*" que implicaba la seudonimización[50] de los datos vinculados[51].

[45] VAN DALEN, Steven - GILDER, Alexander - HOOYDONK, Eric - PONSEN Mark, *System risk indication: An assessment of the Dutch anti-fraud system in the context of data protection and profiling* (Universidad de Utrech, 2016), p. 10. [Visible en: https://n9.cl/yfce6] [Consultado por última vez: 9/08/2021].

[46] Vid. Artículos 64 y 65 de la ley SUWI. [Visible en: https://wetten.overheid.nl/BWBR0013060/2020-01-01/#Hoofdstuk9_Artikel65] [Consultado por última vez: 9/08/2021].

[47] Vid. Decreto SUWI. [Visible en: https://zoek.officielebekendmakingen.nl/stb-2014-320.html] [Consultado por última vez: 9/08/2021].

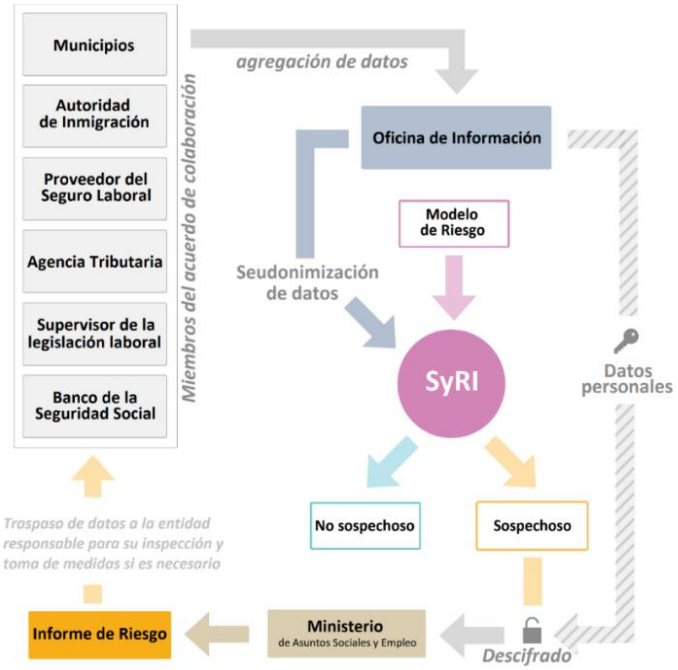
[48] VAN DALEN *et al*, cit. (n. 44), pp. 10-11.

[49] Ibid., p. 11.

[50] La seudonimización, según el artículo 4.5 del RGPD, consiste en "*el tratamiento de datos personales de manera tal que ya no puedan atribuirse a un interesado sin utilizar información adicional, siempre que dicha información adicional figure por separado y esté sujeta a medidas técnicas y organizativas destinadas a garantizar que los datos personales no se atribuyan a una persona física identificada o identificable*". En otras palabras, se trata de un proceso de anonimización reversible.

[51] VAN DALEN *et al*, cit. (n. 44), p.11.

Tabla 1: funcionamiento de SyRI[52]



Del mismo modo, una seguidilla de objeciones, fundadas principalmente en la privacidad de las personas, seguridad de los datos y falta de transparencia, fueron dando paso a la normativa de SyRI. A través de esta normativa, diversas agencias gubernamentales —que en su conjunto forman el paradigma del Estado de Bienestar neerlandés— están autorizadas para formar acuerdos de colaboración, en virtud de los cuales combinan sus bases de datos con el fin de agregar, triangular y analizar dichos datos, dando lugar a informes de riesgo[53].

Un informe de riesgo indica, por su parte, que una persona física es digna de investigación al detectarse un mayor riesgo de posible fraude en el campo de la seguridad social, fiscal y laboral[54]. Los datos susceptibles de ser suministrados constan en una vastísima lista de 17 categorías, los que deben contar con una antigüedad no superior a dos años[55]. Los acuerdos de colaboración, por otro lado, se basan en las prioridades que elaboran anualmente los equipos de intervención del Comité Directivo Nacional, en nombre del Ministro de Asuntos Sociales[56], pues el "responsable del tratamiento" es el Ministerio de Asuntos Sociales[57].

Para el inicio de dicho acuerdo, las instituciones que conforman los equipos deben elevar una solicitud al mencionado ministerio, indicando una serie de aspectos organizativos: instituciones, objetivos, fecha de inicio y duración, necesidad de los datos, método de notificación y modelo

[52] La versión original de este diagrama de flujo que ilustra el funcionamiento de SyRI fue puede consultarse en: BRAUN, Ilja *High-Risk Citizens,* algorithmwatch (2019) [Visible en: https://algorithmwatch.org/en/high-risk-citizens/] [Consultado por última vez: 9/08/2021].
[53] VAN DALEN *et al*, cit. (n. 44), pp. 12- 13.
[54] Artículo 65, párrafo 2 de la Ley SUW.
[55] Artículo 5a.1, párrafo 3 del Decreto SUWI.
[56] VAN DALEN *et al*, cit. (n. 44), p.13.
[57] El RGPD define "responsable del tratamiento" como "persona física o jurídica, autoridad pública, servicio u otro organismo que, solo o junto con otros, determine los fines y medios del tratamiento". Vid. Artículo 4.7 del RGPD.

de riesgo[58]. Además, las instituciones deben velar por los criterios de proporcionalidad y necesariedad[59]. Una vez aprobada la solicitud por el Ministerio, los datos deben ser proporcionados a la Agencia de Inteligencia ("*Inlichtingenbureau*"), quien lleva a cabo el procesamiento[60]. Los datos se seudo nimizan, combinan en perfiles y finalmente son vinculados a través de un modelo de riesgo preestablecido[61].

Cada proyecto SyRI tiene su propio modelo de riesgo, basado en indicadores que sugieren una mayor probabilidad de cometer fraude[62]. Establecido el análisis, el sistema indica las alarmas de riesgo a partir del modelo subyacente. Los resultados obtenidos se clasifican en 2 categorías "sospechoso" y "no sospechoso". En el primer caso, los conjuntos datos de datos se descifran y se crean perfiles de riesgo de personas físicas determinadas, mientras que en el caso de los no sospechosos los datos se eliminan en un período máximo de cuatro semanas[63]. Luego, los perfiles de riesgo elaborados son enviados al ministerio para ser analizados nuevamente.

Finalmente, se agregan los informes a un registro central, donde se almacenan por dos años[64]. Durante este período, los informes están disponibles para los órganos administrativos responsables, que pueden llevar a cabo investigaciones posteriores caso a caso; La Agencia Tributaria, la Inspección de Asuntos Sociales y Empleo, y el Banco de la Seguridad Social disponen de acceso directo a SyRI, mientras que otros órganos como los municipios o la fiscalía deben pedir permiso al ministerio si requieren analizar los informes[65]. Las personas pueden consultar en el mencionado registro si su nombre ha dado lugar a un informe de riesgo por el sistema[66].

En cuanto al despliegue de SyRI, desde que su uso se reguló en la legislación, se realizaron los siguientes proyectos: "GALOP II", "Abordar el fraude Afrikaanderwijk en Rotterdam", "Barrios vulnerables de WGA Capelle aan den IJssel", "WGA Rotterdam Bloemhof y Hillesluis" y "WGA Haarlem Schalkwijk"[67].

En lo que se refiere a los modelos algorítmicos empleados en su despliegue, pese a que existen algunas versiones cruzadas, no existe información pública disponible para su verificabilidad, lo cual nos sitúa en una vertiente especulativa. Incluso desde la óptica del operador jurisdiccional —en el caso NJCM et al contra los Países Bajos— se determinó que no se podía determinar con exactitud qué es SyRI. Esto es una decisión deliberada del Estado, pues sólo se informa que existe un modelo algorítmico, indicadores de riesgo y un punto de corte, los cuales se opta por mantener en secreto para garantizar la eficacia operativa del sistema[68]. La parte demandante, por otro lado, basándose en un informe que la División Asesora del Consejo de Estado presentó al gabinete, señala que se habría hecho uso de "aprendizaje profundo" (en adelante, "DL" por sus

[58] VAN DALEN *et al*, cit. (n. 44), p.13.
[59] Ibid.
[60] Se trata de una parte privada que actúa bajo la responsabilidad del Departamento, de conformidad con el artículo 5a.2, sección 3 y 5a.1, sección 7 "Decreto SUWI", en calidad de "procesador" en el sentido de la Wbp.
[61] VAN DALEN *et al*, cit. (n. 44), p.13.
[62] Decreto SUWI art. 5ª. 1.7.
[63] VAN DALEN *et al*, cit. (n. 44), p.13.
[64] Ibid.
[65] Ibid.
[66] Ibid.
[67] Cit. (n. 3), párrafo 3.10. de la sentencia.
[68] Ibid., párrafo 6.49. de la sentencia.

siglas en inglés)[69]. Finalmente, aunque no pudiese verificarse, el tribunal sostuvo, junto a la mencionada División, que el uso de SyRI "encaja" con sistemas de DL[70].

Cómo es posible advertir, el sistema contaba con numerosas garantías: la seudonimización, el período limitado de uso, requisitos organizativos, etc. Sin embargo, tras un informe de riesgo se encuentra un proceso que no puede ser totalmente comprendido o auditado por los sujetos de datos. Asimismo, las entradas y salidas son, o bien completamente desconocidas, o conocidas sólo parcialmente por los afectados. Y si bien existe un registro accesible para los titulares de los datos, no se proporciona información sobre cómo se ha llegado a una determinada clasificación de riesgo. Tampoco se informa concretamente qué datos personales se han utilizado. La opacidad se extiende también respecto a terceros expertos, pues no se verifican auditorías externas al sistema. Sólo se prevé un control interno que, por lo demás, tampoco transparentaba sus criterios. Bajo estas circunstancias, es comprensible que la opacidad sea un tema central en el debate acerca de la utilización de SyRI. Como se discutirá en la próxima sección, esto figura también como una de las principales inquietudes en torno al uso de perfiles de riesgo automatizados basados en aprendizaje automático.

II. TRANSPARENCIA Y OPACIDAD ALGORÍTMICA EN EL PROCESO DE PERFILADO

Como se adelantaba, la opacidad está en el ojo del huracán en lo que se refiere al debate sobre el uso de ML. Comúnmente se alude a los sistemas algorítmicos como opacos en el sentido de que el receptor de la inferencia de un algoritmo "rara vez se tendrá un sentido concreto de cómo o por qué se ha llegado a una clasificación particular a partir de las entradas"[71]. Para ilustrar eficazmente este punto, la antesala necesaria, es determinar qué se quiere decir con transparencia en el contexto de la elaboración automatizada de perfiles. Con este propósito, las siguientes líneas examinan, a partir de la literatura especializada, este binomio conceptual —transparencia y opacidad— en un sentido relevante para la elaboración de perfiles y la protección de datos.

Procediendo en este orden, el análisis sigue un enfoque vagamente relacional, lo que implica una concepción del fenómeno de la transparencia, no como un fenómeno meramente informacional, sino como una relación entre un agente y un receptor, de acuerdo con una determinada situación comunicativa, distinguiendo así entre distintos factores contextuales[72]. La primera parte aborda lo relativo al contenido informacional de la transparencia. Seguido de ello, se ilustran los valores en competencia que subyacen al principio de transparencia. Y para cerrar esta sección, se aborda el problema de la opacidad algorítmica, distinguiendo tres dimensiones; intencional, alfabética e intrínseca. Cada una se vincula con un presupuesto de la transparencia, a saber: accesibilidad, comprensibilidad del grupo objetivo o alfabetismo técnico, e interpretabilidad y explicabilidad.

[69] Ibid., párrafo 6.46. de la sentencia.
[70] Ibid., párrafo 6.51. de la sentencia.
[71] BURRELL, cit. (n. 28), p. 1.
[72] FELZMANN, Heike - FOSH-VILLARONGA, Eduard - LUTZ, Christoph - LARRIEUX-Tamo, Aurelia*, Robots and transparency: The multiple dimensions of transparency in the context of robot technologies,* en *IEEE Robotics & Automation Magazine* 26 (2019*)* 2, pp. 71-78.

1. El principio de transparencia algorítmica

La transparencia es el principio rector del derecho a la protección de datos personales. No obstante, en el proceso de elaboración automatizada de perfiles la transparencia es siempre un fenómeno complejo, el cual puede observarse de manera distinta en las diversas etapas del ciclo algorítmico vistas con anterioridad. No se trata, por ende, de un fenómeno estático, sino de un proceso abierto a una plétora de variables interactivas y en perpetuo movimiento[73]. Desde esta perspectiva, el estándar óptimo de transparencia se puede considerar "a nivel de todo el modelo, vale decir, a nivel de componentes individuales (por ejemplo, parámetros), a nivel de un algoritmo de entrenamiento particular"[74], así como también respecto de sus iteraciones a lo largo del tiempo. En esta línea, es válido afirmar que "en el sentido más estricto, un modelo será transparente si una persona puede contemplar el modelo completo a la vez"[75]. Sin embargo, este estándar idealista debe contrastarse con una situación fáctica que involucra serie de factores contextuales para producir la "comprensibilidad" del receptor.

Tabla 2[76]

Fases de la Transparencia	Grupos objetivos
A. Recolección de datos	1. Público en general
B. Construcción del modelo	2. Instituciones internas y externas
C. Implementación (subetapas de recopilación, análisis, inferencia y uso *ex post*)	3. Sujetos afectados (directos e indirectos)

Siguiendo esta línea, es posible identificar 2 dimensiones temporales relevantes de la transparencia, según se refiera a la "funcionalidad del sistema" o a una "decisión específica"[77]. La primera se refiere a la "importancia, las consecuencias previstas y la funcionalidad general de un sistema automatizado de toma de decisiones", lo que implicaría revelar, por ejemplo, "la especificación de requisitos del sistema, árboles de decisión, modelos predefinidos, criterios y

[73] Ibid.

[74] LEPRI, Bruno - OLIVER, Nuria - LETOUZÉ, Emmanuel - PENTLAND, Alex - VINCK, Patrick, *Fair, transparent, and accountable algorithmic decision-making processes,* en *Philosophy & Technology* 31 (2018) 4, p. 9.

[75] Ibid.

[76] ZARSKY, Tal, *Transparent Predictions,* en *Revista de Derecho de la Universidad de Illinois* 4 (2013), p. 1533. [Visible en: https://ssrn.com/abstract=2324240] [Consultado por última vez: 9/08/2021].
Esta tabla está basada en la que presenta Zarsky en "*Transparent predictions*", aunque con algunas diferencias. La columna de "etapas de la transparencia" de Zarsky sólo contempla "recopilación, análisis y uso *ex post*", que en este caso se encuentran encapsuladas dentro de la fase de "implementación".

[77] WACHTER, Sandra - MITTELSTADT, Brent - FLORDI, Luciano, *Why a right to explanation of automated decision-making does not exist in the general data protection regulation,* en *International Data Privacy Law* 7 (2017) 2, pp. 78-79. [Visible en: https://doi.org/10.1093/idpl/ipx005] [Consultado por última vez: 9/08/2021].

estructuras de clasificación"[78]. Por ende, esta noción se aplica a todas las fases anteriores a la implementación del modelo. Las decisiones específicas, por otro lado, aluden a "la justificación, las razones y las circunstancias individuales de una decisión automatizada específica", por lo tanto, incluiría "la ponderación de características, reglas de decisión específicas de caso definidas por la máquina, información sobre grupos de referencia o perfiles", entre otros[79]. Es decir, incluye todas las subetapas de la fase de implementación.

En dichos términos, es posible distinguir una "dimensión prospectiva" y una dimensión "retrospectiva" de la transparencia, según el momento al que se oriente (*ex ante* o *ex post* de una decisión específica)[80]. La transparencia prospectiva se requiere *ex ante*, puesto que "informa sobre el procesamiento de datos y el funcionamiento del sistema por adelantado"[81], vale decir, se corresponde con los elementos de la transparencia relativa a la "funcionalidad del sistema"; en fin, describe cómo el sistema toma decisiones en general. La transparencia retrospectiva, por otro lado, se refiere a "explicaciones y fundamentos *post hoc*"[82]. Así, "la transparencia prospectiva es un elemento necesario para obtener rendición de cuentas", mientras que "la transparencia retrospectiva se requiere para fines de auditoría"[83]. Por lo tanto, para que un sistema algorítmico de toma de decisiones tenga transparencia retrospectiva, "uno debería poder inspeccionar sus partes internas, descomponer una decisión para comprender la estructura y el sistema de pesaje dentro del sistema y, en última instancia, explicar una decisión"[84].

Siguiendo esta métrica de dos tipos distintos de transparencia, Edwards y Veale señalan que serían lógicamente posibles 2 vías concretas para explicar las decisiones: las explicaciones centradas en el modelo (*Model-Centric Explanations*, en adelante MCE por sus siglas) y las centradas en el sujeto (*Subject-Centric Explanations*, en adelante SCE)[85]. Las MCE buscan encapsular la funcionalidad general del sistema, mientras que las SCE se crean a partir de un "registro de entrada"[86]. Por ende, las primeras son útiles para la función prospectiva, mientras que las segundas se orientan a la función retrospectiva. Empero, las SCE también "pueden proporcionarse frente a consultas ficticias", por lo que en principio podrían satisfacer la función prospectiva[87].

Así, la MCE podría incluir: "información de configuración" ("intenciones del proceso, la familia del modelo y los parámetros utilizados para especificarlo antes del entrenamiento"); "metadatos de entrenamiento": ("estadísticas resumidas y descripciones cualitativas de los datos de entrada y salida"); "métricas de rendimiento" ("información sobre la capacidad de predicción del modelo en datos no vistos"); "lógicas globales estimadas" ("formas simplificadas, promediadas y comprensibles para el ser humano de cómo las entradas se convierten en salidas")

[78] Ibid.
[79] Ibid.
[80] FELZMANN, Heike - FOSH-VILLARONGA, Eduard - LUTZ, Christoph - LARRIEUX-Tamo, Aurelia, *Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns*, en *Big Data & Society* 6 (2019) 1, pp. 2-3.
[81] Ibid.
[82] Ibid.
[83] Ibid.
[84] Ibid.
[85] EDWARDS - VEALE, cit. (n. 2), pp. 61-65.
[86] Ibid.
[87] Ibid.

y, por último; "información sobre el proceso" ("cómo se ha examinado el modelo para detectar propiedades no deseadas")[88].

En el caso de las SCE, por otro lado, es posible distinguir 4 tipos: "explicaciones centradas en la sensibilidad" ("¿qué cambios en mis datos de entrada habrían hecho que mi decisión resultara diferente?"); "explicaciones centradas en el sujeto" ("¿qué registros de datos utilizados para entrenar este modelo son más similares a los míos?"); "explicaciones demográficas centradas en el sujeto" ("¿cuáles son las características de las personas que recibieron un trato similar al mío?"); y "explicaciones centradas en el rendimiento del sujeto" ("¿qué confianza tiene en mi resultado?")[89].

En los modelos algorítmicos de naturaleza completamente simplista o lineal predefinida (*e.g.* árbol de decisiones), en principio, sería posible que los fundamentos de las decisiones específicas sean revelados *ex ante*[90]. Por lo tanto, "cuando sea posible generar una explicación de la funcionalidad del sistema, debería ser posible generar una explicación de decisiones específicas dados los datos de entrada"[91]. Cuando se tratan modelos complejos, en cambio, la dimensión temporal de la transparencia puede significar "pertinencia futura, revelación anticipada, cierre continuo o visibilidad *post hoc*, en fin, diferentes momentos en el tiempo que se pueden requerir para obtener transparencia"[92]. Esto es relevante, especialmente "para fines de auditoría", pues sin dimensiones temporales se pierde la posibilidad de "ver las iteraciones anteriores, comprender cómo funcionaban, por qué cambiaban y cómo sus componentes interactivos constituían en realidad sistemas diferentes"[93]. Esto, dado el hecho de que los sistemas "cambian con el tiempo, especialmente rápido en el contexto de los sistemas computacionales en red", de modo que, incluso si pudiéramos ver todos los elementos del bucle (*i.e.* código, datos de capacitación, etc) sólo nos "daría una visión instantánea y particular de su funcionalidad"[94].

Ahora bien, dado que la transparencia se orienta a proporcionar alguna clase de conocimiento, las políticas de transparencia deben considerar necesariamente al tipo de grupo objetivo al que se dirigen[95]. En este sentido, los receptores se pueden clasificar en tres grandes grupos: el público en general, grupos expertos y los titulares de los datos[96]. El público en general lógicamente es la categoría más amplia e incluye a los otros dos grupos. Dentro de los grupos expertos podemos trazar una distinción según se trate de instituciones gubernamentales, (como podrían ser agentes del propio Estado, por lo que se trataría de una política de transparencia "intrasistémica") o grupos externos (como alguna corporación privada u ONG)[97]. En el caso de los "sujeto de datos" cabe también hacer una distinción. Se ha planteado que los perfiles elaborados mediante ML solo no identifican a una persona, sino a un grupo, puesto que no se basan en característica individuales, sino que la clasificación se realiza en contraste con los demás individuos del

[88] Ibid.
[89] Ibid.
[90] WACHTER *et al*, cit. (n. 76), p. 79.
[91] POWLES, Julia - SELBST, Andrew, *Meaningful Information and the Right to Explanation*, en *Conference on Fairness, Accountability and Transparency* 7 (2018) 4, p. 239.
[92] ANANNY, Mike - CRAWFORd, Kate, *seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,* en *new media & society* 20 (2018) 3, p. 9.
[93] Ibid.
[94] Ibid.
[95] ZARSKY, cit. (n. 75), pp. 1532-1533. También: DE LAAT, cit. (n. 29), p. 527.
[96] Ibid.
[97] Ibid.

conjunto de datos[98]. De este modo, los datos de un individuo "podrían examinarse detenidamente y utilizarse para el conjunto de entrenamiento, generando un árbol de decisiones o un patrón, aunque nunca afectan realmente al individuo concreto"[99]. Así, podemos distinguir a los afectados directamente por la clasificación y a los afectados indirectamente[100].

2. *Fundamentos para la transparencia*

Habitualmente, la transparencia se nos aparece como valiosa en sí misma; todo acto gubernamental debería ser "por defecto" transparente[101]. No obstante, como señala ZARSKY, si no se tiene a la vista una "noción instrumental" de la transparencia, ésta ineludiblemente cederá ante otros intereses "aparentemente más importantes"[102]. En ese sentido, vale la pena considerar los valores que subyacen a las políticas de transparencia. Escapa del propósito de esta memoria proponer una revisión general y exhaustiva. En lugar de ello, se brinda una sucinta descripción de dos asuntos relevantes en el caso SyRI para contextualizar el debate jurídico que sigue; primero, la conexión entre la transparencia, la confianza y la autonomía, y segundo, la relación entre la transparencia y la equidad.

a) Confianza y autonomía

El más intuitivo llamamiento a que la transparencia sea distribuida lo más extendidamente posible —*id est* accesible al público en general y respecto de todas las fases del proceso— surge de su estrecho vínculo con la democracia[103]. Es posible reforzar este punto señalando que, en una sociedad democrática, la transparencia se considera un pilar fundamental, pues promueve la confianza en las acciones gubernamentales y el control de los ciudadanos sobre el poder institucionalizado. En otras palabras, la transparencia otorga legitimidad al poder, puesto que permite "canalizar, regular y controlar los poderes legítimos"[104]. Por ende, la transparencia permite controlar las arbitrariedades, generando confianza y legitimidad para las acciones gubernamentales.

Pero el control y la confianza, vistos desde la perspectiva del sujeto de datos en el contexto de la elaboración de perfiles de riesgo, pueden entenderse como una extensión de la autonomía individual. La vigilancia estatal, desde la perspectiva del perfilamiento, puede tener importantes consecuencias conductuales, como son los "efectos de enfriamiento"[105]. Después de todo, "se actúa de forma diferente si se sabe que los rastros que se dejan serán procesados"[106]. Así, por ejemplo, se puede advertir en el testimonio que recoge la FNV en su plataforma, acerca de un residente de Hillesluis, quien da cuenta del pánico que habría gatillado en su familia la amenaza de SyRI, llegando incluso a dejar de recibir visitas por temor a que el sistema interpretara, que

98 EDWARDS - VEALE, cit. (n. 2), p. 36.
99 ZARSKY, cit. (n. 75), p. 1532.
100 Sin perjuicio de esto, al hablar de "sujetos de datos" se hará alusión a los afectados directamente por la clasificación, salvo que se indique lo contrario.
101 ZARSKY, cit. (n. 75), p. 1531.
102 Ibid.
103 Ibid.
104 GUTWIRTH, Serge - DE HERT, Paul, *regulating profiling in a democratic constitutional state*, en HILDEBRANDT, Mireille - GUTWIRTH, Serge (editores), *Profiling the European citizen* (Springer, Dordrecht Springer, Dordrecht, 2008), p. 277.
105 Vid. SCHAUER, Frederick, *Fear, risk and the First Amendment: Unraveling the chilling effect* (1978), p. 689 en BÜCHI *et al*, cit. (n.14), p. 4.
106 GUTWIRTH - DE HERT, cit. (n. 103), p. 291.

había un "residente adicional", lo que presuntamente les haría perder sus beneficios sociales[107]. En otras palabras, la preocupación por la falta de transparencia, traducida en la incertidumbre sobre cuándo, cómo o qué datos serán tratados en el análisis de riesgo tiene importantes repercusiones conductuales que pueden erosionar la autonomía individual, llegando a materializarse en un efecto constrictivo respecto a conductas legítimas.

Como se verá más adelante, estos asuntos son de no poca relevancia en la resolución del caso SyRI. De momento, basta con señalar que los demandantes acertadamente argumentan que, dada la opacidad con que se utiliza la herramienta, sería posible deducir un "efecto paralizador" que podría amenazar la funcionalidad operativa del sistema. De hecho, el tribunal tuvo especialmente en cuenta que el despliegue de las técnicas de ML requiere de grandes conjuntos de datos, ya que los algoritmos "aprenden" a hacer clasificaciones sobre la base de ejemplos dados, de modo que, sin la confianza suficiente en el sistema, los ciudadanos estarán menos dispuestos a proporcionar datos o habrá menos apoyo para ello[108].

b) Equidad y no discriminación

Otro de los valores que subyacen a las políticas de transparencia en los dominios del ML es la equidad predictiva. Existe un amplio consenso acerca de lo inequitativas que pueden ser las clasificaciones algorítmicas, incluyendo, por supuesto, las clasificaciones a partir de la aplicación de perfiles de riesgo a personas físicas. En este sentido, una de las principales objeciones que puede plantearse a la opacidad es la posibilidad de obliterar algún tipo de sesgo problemático, discriminación basada en atributos sensibles, correlaciones apofénicas o espurias, o bien algún efecto de retroalimentación negativa inadvertido que lleve a un mayor escrutinio a grupos históricamente desfavorecidos. Por tanto, la relevancia de la transparencia es proporcionar las condiciones de verificabilidad de las buenas prácticas en el tratamiento de datos. He ahí el estrecho vínculo del principio de transparencia con la interdicción de arbitrariedad. Esta consideración es especialmente relevante para nuestro caso de estudio, precisamente por la forma en que se despliega SyRI, pues "tiene un efecto discriminatorio y estigmatizador" al focalizar su análisis en "barrios problemáticos", dado que "contribuye a los estereotipos y refuerza una imagen negativa de los vecinos que viven en ellos"[109].

En la elaboración automatizada de perfiles, la discriminación algorítmica puede ocurrir tanto explícita como implícitamente, lo que nos lleva a hablar de discriminación directa e indirecta. Aunque suele ser excepcional, la discriminación directa tiene lugar, en sentido formal, cuando se usan variables sensibles (*e.g.* atributos como la raza, género u orientación sexual) como factor actuarial y además la decisión se ha basado en ellas.

---

[107] "Ik denk dat veel mensen onderschatten wat de dreiging van SyRI met een mens kan doen. Mijn moeder was erg geschrokken, omdat ze een extra logeerbed in huis heeft voor mijn neefjes en nichtjes wanneer ze bij oma logeren. Door de dreiging van SyRI raakte ze in paniek dat de sociale recherche dit zou aanmerken als 'extra inwoner' waardoor ze mogelijk haar toeslagen zou verliezen (…)". (Creo que mucha gente subestima lo que la amenaza de SyRI puede hacerle a una persona. Mi madre estaba muy sorprendida, porque tiene una cama de invitados adicional en la casa para mis sobrinos y sobrinas cuando se quedan con la abuela. La amenaza de SyRI la hizo entrar en pánico de que el departamento de investigación social lo clasificara como un "residente adicional", lo que posiblemente le haría perder sus beneficios.) Heerekop, Annika, "Comunicado de prensa, *FNV en bewoners Rotterdamse wijken Hillesluis en Bloemhof vieren intrekking SyRI-project Rotterdam"*, 16 de julio de 2019. [Visible en: https://n9.cl/ztplkx] [Consultado por última vez: 9/08/2021].

[108] Cit. (n. 3), párrafo 6.5. de la sentencia.

[109] Ibid., párrafo 6,92 de la sentencia.

La discriminación indirecta, por otro lado, se produce cuando a partir de criterios aparentemente neutros clasifican a grupos de personas pertenecientes a una categoría protegida en desventaja injustificada con respecto a otras personas que no pertenecen a dicho grupo, sin utilizar esa categoría como una variable explícita. Este segundo tipo de discriminación, como se desarrolla a continuación, se vincula estrechamente con el problema del sesgo algorítmico[110].

Un primer problema surge, como ha sido registrado por BAROCAS y SELBST, por la posibilidad de que los desarrolladores, aún de manera inconsciente, traspasen sus sesgos en el proceso de localizar relaciones estadísticas en un conjunto de datos[111]. En la misma línea, también existe un riesgo de que los conjuntos de datos de capacitación sobrerrepresenten a ciertos grupos de personas que comparten los mismos atributos —especialmente protegidos— o infrarrepresenten a otros, lo que eventualmente puede traducirse en un impacto dispar[112].

Del mismo modo, se ha registrado cómo el hecho de privilegiar los errores de falsos positivos sobre los falsos negativos, o viceversa, también puede inducir situaciones dispares, llevando desproporcionadamente a un mayor escrutinio a unos por sobre otros[113].

Otra coyuntura surge por el entrenamiento de los algoritmos con datos contaminados con casos de discriminación histórica. El algoritmo los asumirá como ejemplos legítimos en el proceso de aprendizaje, reproduciendo los patrones de discriminación en los resultados de la clasificación[114]. Lo anterior resulta especialmente preocupante, pues se ha registrado cómo los algoritmos pueden no solo reproducir, sino además amplificar las desigualdades escritas en los datos de manera inadvertida[115].

Ciertamente, el más intuitivo resguardo para impedir resultados contaminados con casos de sesgo o discriminación sería, lógicamente, impedir el uso de características sensibles dentro de los conjuntos de datos. Pero esta medida sólo impide la discriminación directa. En el caso de la indirecta, el problema suele ser mucho más profundo puesto que, pese a la eliminación explícita del atributo protegido, el algoritmo aún podría identificar automáticamente *proxys* (o sustitutos de características) correlacionados con el atributo eliminado (*e.g.* en barrios con alta densidad de población migrante, el código postal sueñe operar como un proxy de la raza)[116]. En este contexto, GOODMAN señala que, aún si se lograran localizar todas las correlaciones de algún sustituto problemático, su eliminación puede ser inviable, toda vez que dichas correlaciones pueden

110 Ha de precisarse, sin embargo, que el sesgo no equivale necesariamente a discriminación en sentido jurídico. En términos neutros, un sesgo algorítmico puede ser entendido como una mera desviación estadística cuyo valor ético no viene predeterminado. A veces puede representar, de hecho, una compensación a resultados problemáticos. Por otro lado, están los sesgos eticamente problemáticos. En consecuencia, me referiré a ellos en este último sentido. Vid. LAZCOZ, Guillermo, cit. (43).

111 Vid. BAROCAS - SELBST, Andrew, *Big data's disparate impact,* en *California Law Review* 104 (2016) 3.

112 Ibid.

113 Vid. CHOULDECHOVA, Alexandra, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, en *Big Data* 5 (2017) 2, pp. 153-163.

114 Vid. SCHERMER, Bart, *The limits of privacy in automated profiling and data mining*, en *Computer Law & Security Review* 27 (2011) 1, pp. 45-52.

115 Vid. O`NEIL, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (2016) Nueva York, Crown.

116 Vid. ŻLIOBAITė, Indrė - CUSTERS, Bart, *Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models* en *Artificial Intelligence and Law* 24 (2016) 2, pp. 183-201.

contener información estadística valiosa para lo que se intenta predecir, de modo que su eliminación podría implicar dejar inutilizable al algoritmo[117].

Si no se puede suponer que los sistemas algorítmicos sean justos e imparciales, es ineludible que se requiera alguna forma de "abrir la caja negra" para justificar las decisiones que se tomen a partir de las inferencias. Nótese que con esto no se quiere decir que transparencia sea equivalente a equidad predictiva. El rol de la transparencia es abrir el camino para verificar que los resultados sean equitativos, para detectar casos en que los patrones utilizados puedan ser constitutivos de algún tipo de error, sesgo o discriminación en sentido problemático, o bien para supervisar que se han tomado las medidas orientadas a mitigar los casos injustos[118]. Esta premisa puede desglosarse en varios razonamientos subyacentes.

En primer lugar, con el objeto de asegurar que no se ha empleado directamente como factor actuarial dentro del análisis información sensible o discriminatoria, debe revelarse qué tipo de datos se emplean en el análisis. Asimismo, para saber si la clasificación se ha basado en algún atributo sensible, se requiere algún conocimiento sobre las causas últimas en que se fundamenta una clasificación (*i.e.* en qué criterios se ha basado la inferencia), así como también para la verificación de que no están operando *proxys* que puedan resultar problemáticas.

En este sentido, desde la perspectiva del sujeto adversamente afectado por el resultado, la transparencia retrospectiva aparece como una pieza clave para comprender cómo y por qué se le ha clasificado como tal, y eventualmente impugnar una decisión injusta. Sin embargo, tal como señala GOODMAN, puede que las salvaguardas resulten insuficientes si no se acompañan del rol técnico y detallado que solo podrían proporcionar los grupos expertos, destacando la prioridad de la realización de auditorías y evaluaciones de impacto de protección de datos (en adelante, EIPD) que evalúen los algoritmos caso por caso, donde se requeriría un flujo de información mucho más profundo[119].

En consecuencia, frente al problema de la discriminación y la equidad, tanto la perspectiva ex ante como ex posteriori de la transparencia son piezas fundamentales, y aparecen íntimamente ligadas al ideal de control efectivo en la toma de decisiones algorítmicas, respecto tanto del sujeto de datos como de grupos expertos que cumplan tareas de auditoría y supervisión.

3. *Opacidad algorítmica*

Sin perjuicio de lo anterior, las políticas de transparencia habitualmente sólo podrán dar lugar a una exposición limitada de las etapas del ciclo algorítmico, ya sea por la conveniencia de la reserva para la funcionalidad del sistema, por la carencia de habilidades técnicas para entender la jerga computacional del receptor o por la mera complejidad del modelo subyacente. Este fenómeno es lo que se conoce como "opacidad algorítmica". Siguiendo a BURRELL, se trata de un concepto polisémico, por lo que un marco útil para atender a sus dimensiones es distinguir entre tres tipos distintos de opacidad, según se trate de: opacidad intencional, alfabética e intrínseca[120]. *A contrario sensu*, cada una de estas dimensiones puede ser reconducida a un presupuesto de la transparencia, a saber: accesibilidad, comprensibilidad e interpretabilidad. A continuación, se procede a analizar las categorías presentadas por BURRELL, encapsulando las

[117] GOODMAN, cit. (n. 42), pp. 2-4.
[118] ZARSKY, cit. (n. 75), p. 1520.
[119] GOODMAN, cit. (n. 42), pp. 3-6.
[120] BURRELL, cit. (n. 28), p. 3.

categorías alfabética e intrínseca como parte de un conjunto más grande, la dimensión no intencional de la opacidad.

a) Opacidad intencional

La opacidad intencional alude a la elección deliberada de no divulgar ciertos aspectos del proceso algorítmico[121]. Por ende, se vincula con la accesibilidad de la información. Bajo esta categoría podemos identificar 3 argumentos que típicamente suelen argüirse para restringir las políticas de transparencia: la privacidad, la ventaja competitiva de las empresas y el juego estratégico[122].

El primero de ellos plantea que la revelación de las entradas y salidas del algoritmo podría afectar negativamente el derecho a la privacidad de terceros[123]. Este planteamiento es especialmente relevante en aquellos casos en que se emplea información sensible, tal como la raza, la religión y la etnia. Piénsese, por ejemplo, en el análisis desplegado por SyRI, el cual se emplea habitualmente en barrios con altos índices de pobreza y vulnerabilidad. Como subraya también ZARSKY, al permitir que terceros accedan a esta información, se podría estigmatizar a ciertos individuos por pertenecer a ciertos grupos caracterizados como vulnerables[124]. Además, De Laat argumenta que, dado que "en la era del Big Data, los datos se distribuyen o venden de forma rutinaria a terceros", la accesibilidad supone un riesgo para la seguridad de los datos personales[125].

Otro posible fundamento de este tipo de opacidad resulta cuando un Estado ha contratado con un privado para el uso del programa, donde puede existir una cláusula de confidencialidad, sustentada por el interés comercial legítimo de mantener secreto el código fuente, sea en virtud de la protección de la propiedad intelectual o la devaluación del valor comercial[126]. Esta postura plantea que la publicidad del algoritmo afectaría el derecho de propiedad intelectual y terminaría por desincentivar la ventaja competitiva. La propiedad, en este sentido, permite a los desarrolladores del algoritmo mantener una ventaja respecto de los demás competidores[127]. Si bien esta objeción podría pensarse que es impropia del sector público, eso no es precisamente cierto, pues es cada vez más común que las administraciones compren solamente las licencias de uso en lugar del software[128]. Por lo tanto, el alcance de esta objeción también es aplicable a nuestro caso de estudio.

Como señala ZARSKY, el contraargumento más común a las políticas de transparencia es que podría socavar el "el objetivo que el modelo de predicción pretende alcanzar"[129]. Es decir, la opacidad constituye un medio necesario para "evitar el juego estratégico", donde puede que sea necesario mantener el secreto del código, las categorías de entrada, las ponderaciones realizadas y las categorías de salida[130]. Por este motivo también suele hablarse de "opacidad estratégica". El juego estratégico (o *gaming the algorithm*) consiste en una conducta del sujeto de datos en virtud

121 Ibid.
122 DE LAAT, cit. (n. 29), pp. 533-534.
123 BURRELL, cit. (n. 28), pp. 3-4.
124 ZARSKY, cit. (n. 75), pp. 1560 y ss.
125 DE LAAT, cit. (n. 29), pp. 535-536.
126 Ibid.
127 Ibid.
128 Ibid.
129 ZARSKY, cit. (n. 75), p. 1553.
130 LEPRI *et al*, cit. (n. 73), p 10.

de la cual se busca alterar el resultado de la clasificación del algoritmo, sin alterar la "característica clave" que se intenta predecir a través de él[131].

Así, por ejemplo, en nuestro caso de estudio, la "característica clave" sería el comportamiento fraudulento en materia de asistencia y seguridad social. De este modo, tenemos que distinguir dos escenarios posibles. Si el sujeto modifica su conducta y dejase de cometer fraude o no lo cometiese, dando como resultado de clasificación que el individuo tiene baja probabilidad de cometer fraude, dicho cambio de conducta alteraría la característica clave y, por lo tanto, no es el caso que exista juego estratégico, pues la predicción seguiría siendo correcta. En cambio, si el individuo infractor manipula su comportamiento con el fin de cambiar las variables sustitutivas empleadas para detectar el fraude, pero persistiera en su conducta fraudulenta, no se estaría alterando el factor clave, por lo que sí es el caso de que hubiese juego con el algoritmo.

En el caso de SyRI, como se verá en profundidad más adelante, el Estado precisamente apuntó a que el funcionamiento del algoritmo debía ser oscuro, o de lo contrario los individuos podrían jugar con el sistema, socavando toda posibilidad de funcionalidad operativa. En síntesis, este argumento consiste en que los potenciales defraudadores identificarán las señales de alerta en que se basa el algoritmo y podrán "jugar con el sistema", por lo que socavaría su finalidad.

Por último, cabe mencionar a lo que STOHL *et al.* denominan "opacidad inadvertida", donde los obligados en materia de transparencia proporcionan deliberadamente tanta información que las partes relevantes se "ocultan inadvertidamente", puesto que los receptores "tardarán tanto tiempo y esfuerzo en tamizar" la información que "se distraerán de la información central que el agente desea ocultar"[132].

b) Opacidad no intencional

Además de ser deliberada, la opacidad puede producirse por la falta de "comprensibilidad" de los receptores. Esta premisa nos lleva a la segunda dimensión de la opacidad, *id est* el "analfabetismo técnico"[133]. Con esto se alude a la "carencia de las habilidades técnicas" para entender el lenguaje y los modelos computacionales basados en datos[134]. Esta noción se vincula al nivel educativo de los grupos objetivos, sus conocimientos en computación, familiaridad con otros procesos automatizados, entre otros[135].

Por último, está el problema de la opacidad intrínseca, vinculado a la interpretabilidad de los modelos[136]. Se trata de un tipo de opacidad que alcanza incluso a los propios desarrolladores, quienes pueden "ser incapaces de explicar cómo funciona un sistema de estas características", vale decir, "qué partes son esenciales para su funcionamiento o cómo la naturaleza efímera de las representaciones computacionales es compatible con las leyes de transparencia"[137]. En tales casos, aunque sea posible "ver el interior del sistema" (acceso a los datos de entrada, el código fuente, los datos de entrenamiento, etc.), puede que ello no implique necesariamente entender

131 BAMBAUER, Jane - ZARSKY, Tal, *The algorithm game,* en *Notre Dame L. Rev.* 94 (2018) 1, pp. 6-11.
132 STOHL, Cynthia - STOHL, Michael - LEONARDI, Paul, *Managing opacity: Information visibility and the paradox of transparency in the digital age*, en *the digital age. International Journal of Communication Systems International Journal of Communication* 10 (2016) 15, pp. 133-134 en ANANNY - CRAWFORd, cit. (n. 91), p. 7.
133 BURRELL, cit. (n. 28), p. 4.
134 LEPRI *et al*, cit. (n. 73), p 9.
135 Ibid.
136 BURRELL, cit. (n. 28), pp. 4-5.
137 ANANNY - CRAWFORd, cit. (n. 91), pp. 9-10.

cómo se ha tomado una decisión[138]. Se alude, por ende, a la naturaleza de ciertos dominios, como el DL, donde se produce un desajuste entre los "procedimientos matemáticos" y los "estilos de interpretación semántica"[139].

En síntesis, en la literatura sobre ML las dimensiones de la opacidad puedes ser encausadas por diversas razones. Todas ellas son reconducibles a segmentos específicos de la transparencia (*i.e.* accesibilidad, comprensibilidad e interpretabilidad del modelo), y referidas a diferentes grupos de la sociedad. Lo anterior sin perjuicio de que estas capas de opacidad puedan existir, como a menudo sucede, simultáneamente.

III. TRANSPARENCIA Y PERFILES EN EL SISTEMA EUROPEO DE PROTECCIÓN DE DATOS

El principio de transparencia constituye la piedra angular del derecho a la protección de datos. Históricamente, la protección de los datos personales se ha abordado de manera genérica como un aspecto subordinado al derecho al respeto de la vida privada. Sin embargo, como resultado lógico de la proliferación de tecnologías basadas en macrodatos y su creciente valor —económico y social—, se ha reconocido la importancia de considerar este derecho de forma autónoma.

En el contexto de la Unión Europea (UE), esta distinción se refleja en los artículos 7 y 8 de la Carta de los Derechos Fundamentales de la Unión Europea (en adelante, Carta de la UE), donde el primero garantiza el derecho a la privacidad y el segundo el derecho a la protección de datos personales. En relación a este último, el RGPD, que ha sido directamente aplicable a los Estados de la UE desde 2018, establece una regulación exhaustiva. Dado el alcance limitado de este documento, nos centraremos en el análisis de la regulación relacionada con la elaboración automatizada de perfiles y el régimen de transparencia establecido por dicho instrumento.

El examen toma como punto de partida el artículo 22 sobre las decisiones únicamente automatizadas y el considerando 71, en lo que se refiere al derecho a la explicación. Seguido de ello, se analizan conjuntamente los derechos de notificación y acceso contenidos en los artículos 13 a 15, a propósito del derecho a acceder a la lógica interna. Y para concluir, se abordarán someramente las herramientas orientadas al control experto.

1. El derecho a la protección de datos personales en el derecho de la Unión Europea

El principio de transparencia, aunque no se encuentra explicitado en la Carta de la UE —a diferencia de lo que ocurre con el RGPD— se encuentra inexorablemente ligado a la naturaleza jurídica del derecho a la protección de datos personales. Por consiguiente, las preocupaciones sobre la opacidad algorítmica en la elaboración de perfiles, cuando se tratan datos de carácter personal, abarcarán un enorme peso en materia de protección de datos personales.

Como se indicaba, resulta crucial para atender a este vínculo el contraste que se genera entre la protección de datos y el derecho a la privacidad. Pues si bien históricamente recibieron un tratamiento indistintivo, lo cierto es que apuntan a dimensiones muy distintas. Así lo postularon hace más de un decenio GUTWIRTH y DE HERT, quienes identificaron que mientras la privacidad se vincula con las "herramientas que tienden a garantizar la no injerencia en asuntos individuales

[138] Ibid.
[139] BURRELL, cit. (n. 28), pp. 2-10.

y la opacidad del individuo", el derecho a la protección de datos personales, por otro lado, "se orienta a la transparencia y la responsabilidad"[140]. Esto no obsta, claro, que a menudo se encuentren regulaciones superpuestas, pudiendo hallarse herramientas de opacidad en la regulación de protección de datos, (*e.g.* el principio de limitación de finalidad en la regulación sobre protección de datos), así como herramientas de transparencia en el derecho a la privacidad (*e.g.* las condiciones de intervención del derecho a la privacidad)[141].

En el plano de la Unión Europea, el derecho a la privacidad se encuentra consagrado en el artículo 7 de la Carta de la UE, que a su vez reproduce el texto del artículo 8 del Convenio Europeo para la CEDH, el cual dispone: "1. *Toda persona tiene derecho a que se respete su vida privada y familiar, su domicilio y su correspondencia*". No se trata, sin embargo, de un derecho absoluto. Existen condiciones para una injerencia justificada, las cuales se explicitan en apartado 2 del mencionado artículo. De acuerdo con el inciso segundo de la disposición referida, las interferencias estarán justificadas si: a) están de acuerdo con la ley; b) persiguen un objetivo legítimo; y c) son necesarias en una sociedad democrática[142].

Para que una injerencia esté de acuerdo con la ley debe tener un fundamento en el derecho interno, dicha normativa debe ser accesible, sus efectos previsibles y su uso compatible con el Estado de Derecho; será accesible si está disponible públicamente, sus efectos previsibles si su formulación es lo suficientemente precisa como para que un individuo pueda regular su conducta en consecuencia y su uso compatible con el Estado de Derecho si impide interferencias arbitrarias en los derechos de un individuo[143]. Los objetivos que se consideran legítimos son: *"los intereses de la seguridad nacional, la seguridad pública o el bienestar económico del país, la prevención de desórdenes o delitos, la protección de la salud o la moral, o la protección de los derechos y libertades de los demás"*[144]. Y, por último, una injerencia está justificada si es necesaria en una sociedad democrática, es decir, si coincide con una "necesidad social acuciante" y es la "proporcional al objetivo legítimo que se persigue"[145].

El derecho a la protección de datos personales, por otra parte, encuentra su reconocimiento expreso en el artículo 8 de la Carta de la UE, que dispone: "1. *Toda persona tiene derecho a la protección de los datos de carácter personal que la conciernan*"[146]. Esta disposición, además de hacer la distinción con el derecho a la privacidad, también establece algunas garantías específicas en los párrafos 2 y 3, a saber: que "*los datos personales se deben tratar en modo leal, para fines concretos y sobre la base del consentimiento de la persona afectada o en virtud de otro fundamento legítimo previsto por la ley*"; que "*toda persona tiene derecho a acceder a los datos recopilados que la conciernan y a su derecho a su rectificación*"; y en

---

140 GUTWIRTH - DE HERT, cit. (n. 103), pp. 276-278.

141 Ibid.

142 "*No podrá haber injerencia de la autoridad pública en el ejercicio de este derecho sino en tanto en cuanto esta injerencia esté prevista por la ley y constituya una medida que, en una sociedad democrática, sea necesaria para la seguridad nacional, la seguridad pública, el bienestar económico del país, la defensa del orden y la prevención de las infracciones penales, la protección de la salud o de la moral, o la protección de los derechos y las libertades de los demás.*" Art. 8 CEDH, apartado 2.

143 GREER, Steven, *The exceptions to Articles 8 to 11 of the European Convention on Human Rights*, en *Human rights files No 15 Council of Europe Publishing* (1997), pp. 9-14.

144 Art. 8.2. del CEDH.

145 GREER, cit. (n. 142), pp. 9-14.

146 Pese a que no se regula de manera autónoma en el CEDH, se ha entendido que su art. 8 contiene el mismo alance que el artículo 8. Así lo señala el art. 52 de la Carta ("*sobre la interpretación de los derechos y principios*"): "3) *En la medida en que la presente Carta contenga derechos que correspondan a derechos garantizados por el Convenio Europeo para la Protección de los Derechos Humanos y de las Libertades Fundamentales, su sentido y alcance serán iguales a los que les confiere dicho Convenio. Esta disposición no obstará a que el Derecho de la Unión conceda una protección más extensa*."

el inciso 3 dispone que "*el respeto de estas normas quedará sujeto al control de una autoridad independiente*". Asimismo, este derecho tiene su desarrollo en el RGPD, cuya elucidación se presenta a continuación.

2. El principio de transparencia en el Reglamento General de Protección de Datos

Desde el 24 de mayo de 2016, el RGPD ha entrado en vigor, y es directamente aplicable a partir del 25 de mayo de 2018, por lo que la antigua DPD se encuentra actualmente derogada. En virtud de su art. 5, apartado 1, la transparencia se abrió paso, por primera vez expresamente, como principio jurídico vinculante de la protección de datos. Esta norma dispone que los datos deben ser "*procesados de manera legal, justa y transparente*"[147]. De este modo, la norma vincula la noción de transparencia a la "legalidad y equidad"[148].

El requisito de legalidad no ha sido añadido por mera costumbre retórica, sino que se conecta con las bases jurídicas establecidas en el artículo 6 para un tratamiento de datos lícito. Para nuestro caso de estudio, especial relevancia cobra que el tratamiento se requiera "*para cumplir con una obligación legal a la que está sujeto el controlador de datos*" (art 6. lit c.), y que sea "*necesario para el cumplimiento de una tarea de interés público o de una tarea en el contexto del ejercicio de la autoridad que se ha encomendado al controlador de datos*" (art. 6 lit. e.)[149]. Aunque no únicamente, estas causales parecen ser las justificaciones más adecuadas para la elaboración automatizada de perfiles de riesgo de fraude en el sector público.

El requisito de equidad, por otro lado, suele estar vinculado a la interdicción de la arbitrariedad y el sesgo algorítmico. En esta línea, el párrafo 2 del considerando 71 detalla algunas medidas para *"asegurar un procesamiento justo y transparente"*, entre las cuales se incluyen aquellos *"procedimientos matemáticos o estadísticos apropiados para la elaboración de perfiles*", así como "*medidas técnicas y organizativas"* para "*prevenir, entre otras cosas, los efectos discriminatorios sobre las personas físicas por motivos de origen racial o étnico, opiniones políticas, religión o creencias, afiliación sindical, estado genético o de salud u orientación sexual*" (es decir, categorías protegidas de datos). De ahí que la transparencia esté íntimamente imbricada con la verificabilidad de la equidad en el tratamiento de datos personales.

La transparencia se debe vincular, además, con la comprensibilidad y accesibilidad de la información que debe proporcionarse a los receptores. Así, el considerando 58 establece que "*el principio de transparencia exige que toda información dirigida al público o al interesado sea concisa, fácilmente accesible y fácil de entender, y que se utilice un lenguaje claro y sencillo, y, además, en su caso, se visualice*", fórmula que se replica en otros considerandos, así como en el artículo 12 del RGPD dentro del texto vinculante[150].

---

147 Art. 5, apartado 1 del RGPD.

148 FELZMANN *et al,* cit. (n. 79), p. 2.

149 El Reglamento establece que el fundamento jurídico del apartado f) no se aplica a la tramitación por parte de las instituciones gubernamentales en el contexto del ejercicio de sus funciones. Tampoco son aplicables los fundamentos jurídicos para la tramitación a que se refieren los apartados a), b) y d), por lo que, en principio, sólo serían aplicables al tratamiento con fines evaluar el riesgo de fraude en el sector público las hipótesis mencionadas en c) y e). Vid. artículo 6, apartado 2 del RGPD.

150 "*El responsable del tratamiento tomará las medidas oportunas para facilitar al interesado toda información indicada en los artículos 13 y 14, así como cualquier comunicación con arreglo a los artículos 15 a 22 y 34 relativa al tratamiento, en forma concisa, transparente, inteligible y de fácil acceso, con un lenguaje claro y sencillo."* Artículo 12, apartado 1 del RGPD.

Además, cabe mencionar que, al tratarse de una disposición general, en caso de que no sea posible aplicar normas más específicas para un determinado contexto, es válido recurrir al principio de transparencia como última ratio. No obstante, las manifestaciones de carácter más concreto se encuentran principalmente en dos niveles: por una parte, en los derechos individuales de transparencia de los sujetos de datos; por otra, en las medidas técnicas y organizativas orientadas al control de grupos expertos. A continuación, se procede con el análisis de dichas manifestaciones.

a) Derecho a no ser objeto a decisiones automatizadas

El artículo 22, en su encabezado, establece que *"todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar"*[151]. Esta redacción, muy similar al artículo 15 de la DPD, ha suscitado importantes discusiones que se proceden a desentrañar.

Una primera cuestión surge respecto de si se trata de un derecho de oposición, tal como se consideró históricamente en la DPD, o de una prohibición general[152]. Sin embargo, el conflicto puede considerarse superado a estas alturas, ya que las directrices del Grupo de Trabajo del Artículo 29 (en adelante, "GT 29") aclaran que el artículo "establece una prohibición general para la toma de decisiones basada únicamente en el procesamiento automatizado", por lo que "esta prohibición se aplica independientemente de que el interesado tome o no una acción con respecto al procesamiento de sus datos personales"[153].

En segundo lugar, este derecho distingue entre aquella decisión basada únicamente en tratamiento automatizado y la decisión que no está basada únicamente en dicho tratamiento. A partir de esto se ha planteado que este derecho puede eludirse fácilmente al incorporar la participación nominal de un humano en el bucle algorítmico[154]. Empero, la guía de las directrices del GT 29 aclara que se requiere que la participación sea lo suficientemente significativa para alterar el curso de la clasificación y además "debe tener en cuenta todos los datos pertinentes"[155]. Esta consideración puede ser muy difusa, pues se debe distinguir el grado de autoridad de quien interviene para determinar si la participación ha sido realmente significativa, de acuerdo con los criterios de competencia y sustancia que señala el GT 29, aunque eventualmente podrían considerarse otros factores relevantes de acuerdo con un caso en específico.

---

[151] Un antecedente clave de este artículo es el artículo 15 de la derogada DPD que establecía: "*toda persona tiene derecho a no ser objeto de una decisión que produzca efectos jurídicos sobre ella o que la afecte de manera significativa y que se base únicamente en un tratamiento automatizado de datos destinado a evaluar determinados aspectos personales relacionados con ella, como su rendimiento laboral, su solvencia, su fiabilidad, su conducta, etc*." Luego, en el apartado 2 del artículo se establecen las causales de excepción: "*no obstante, una persona puede ser objeto de una decisión individual automatizada si dicha decisión se adopta: a) en el marco de la celebración o la ejecución de un contrato, siempre que la solicitud de celebración o de ejecución del contrato, presentada por el interesado, haya sido satisfecha o que existan medidas adecuadas para salvaguardar sus intereses legítimos, tales como disposiciones que le permitan exponer su punto de vista o; b) esté autorizada por una ley que también establezca medidas para salvaguardar los intereses legítimos del interesado*".

[152] WACHTER *et al*, cit. (n. 76), p. 94.

[153] Grupo de trabajo sobre protección de datos del artículo 29, "*Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679*" (2018) p. 21. [Visible en: https://www.aepd.es/sites/default/files/2019-12/wp251rev01-es.pdf] [Consultado por última vez: 9/08/2021].

[154] WACHTER *et al*, cit. (n. 76) p. 91-93.

[155] Grupo de trabajo sobre protección de datos del artículo 29, cit. (n. 152), p. 23.

Asimismo, se distingue aquella decisión con efectos jurídicos o una afectación significativa similar de las que tienen poco o ningún efecto. En el caso de un perfil, debemos distinguir primeramente si el perfil se realiza, como vimos, a partir de un tratamiento automatizado de datos personales y sobre personas físicas, en el sentido de la definición que proporciona el RGPD en su artículo 4, apartado 4. Luego, hay que distinguir si la clasificación produce algún efecto, y si el salto inductivo desde la clasificación al efecto mencionado ha estado mediado por participación humana. Por último, lógicamente, hay que evaluar si existe un efecto jurídico o una afectación significativa similar, pues el tratamiento con efecto no significativo sobre las personas físicas no se incluye dentro del ámbito de aplicación de este artículo. Un ejemplo de efecto jurídico, según indica el GT 29, sería "la denegación de una prestación concedida por la ley, como la prestación por hijos o la ayuda a la vivienda", mientras que un efecto similar puede considerarse "afectar significativamente a las circunstancias, al comportamiento o a las elecciones de las personas afectadas" o "tener un impacto prolongado o permanente en el interesado"[156].

Seguido de ello, el apartado 2 del artículo enumera 3 excepciones a este derecho, según si la decisión es: "a) *necesaria para la celebración o la ejecución de un contrato entre el interesado y un responsable del tratamiento*"; "b) *autorizada por el Derecho de la Unión o de los Estados miembros que se aplique al responsable del tratamiento y que establezca asimismo medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado*"; "c) *basada en el consentimiento explícito del interesado"*[157]. Asimismo, el apartado 3 establece que en las excepciones a) y c) el responsable del tratamiento también debe "*adoptar medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado*", entre ellas "*como mínimo el derecho a obtener intervención humana por parte del responsable, a expresar su punto de vista y a impugnar la decisión*".

Por tanto, los requisitos mínimos que deben cumplirse para que la toma de decisiones, que incluye la elaboración de perfiles automatizada, sea legal son: a) que sea necesario para la celebración o ejecución de un contrato; b) la existencia de una habilitación legal que cree una excepción; c) consentimiento explícito, y estableciendo en todos los casos, salvaguardas necesarias para proteger los derechos e intereses legítimos del afectado. Ciertamente, nuestro caso de estudio puede perfectamente subsumirse en la hipótesis "b)". Así lo corrobora el considerando 71, que explícitamente nos indica que sería aplicable "*incluso con fines de control y prevención del fraude y la evasión fiscal*".

Es preciso subrayar que, si se sigue la interpretación del encabezado como una "prohibición general" y no como un mero "derecho de oposición", entonces la existencia de la protección debe tener lugar antes de la decisión automatizada[158]. Es decir, la elaboración automatizada de perfiles no puede tener lugar sin antes haber cumplido con estos requisitos. Sin embargo, la idea de "salvaguardas adecuadas" sigue siendo un concepto "jurídico indeterminado", dado que "se hace en previsión de ulteriores desarrollos nacionales", sin perjuicio de que "no puede comprenderse que dichos desarrollos nacionales no precisen el contenido del término jurídico indeterminado correspondiente al apartado del RGPD que desarrollan, o en el que pretenden apoyarse"[159].

---

156 Ibid., p. 23 y 24.

157 Artículo 22, apartado 2, del RGPD.

158 WACHTER *et al*, cit. (n. 76), p. 94.

159 PARRILLA, Castillo, *La Elaboración de Perfiles Políticos en España tras la Sentencia del Tribunal Constitucional de 22 de mayo de 2019 por la que se Declara Inconstitucional el Art. 58 bis.1 de la Ley Orgánica del Régimen Electoral General - Perfilado político*

En consonancia con esto, se ha generado un acalorado debate acerca de si tales garantías deben o no incluir necesariamente un derecho a la explicación. La coyuntura se genera a partir de un breve documento presentado por GOODMAN y FLAXMAN, en el que afirmaban que el RGPD contenía el "derecho a una explicación" de la toma de decisiones algorítmicas, basándose en la redacción del considerando 71[160]. Este considerando menciona entre las salvaguardas adecuadas para la elaboración automatizada de perfiles "*incluso con fines de control y prevención del fraude y la evasión fiscal*" que se debe incluir necesariamente "*el derecho* (…) *a obtener una explicación de la decisión tomada",* aludiendo, ciertamente, al ámbito de aplicación del artículo 22[161]. Este supuesto, claro, se aduce como base para construir un tipo de transparencia retrospectiva dirigida al individuo afectado. Con todo, nos encontramos con el problema del estatus no vinculante de los considerandos y de que la historia legislativa del artículo 22 parece dar cuenta de que la omisión del "derecho a una explicación" en el texto vinculante fue una omisión deliberada[162].

Por lo demás, si aceptamos la obligatoriedad del derecho a la explicación, aún hay una serie de obstáculos que superar. Para nuestro caso de estudio, por ejemplo, el resultado de la clasificación de riesgo en el caso de SyRI, ¿constituye una decisión únicamente automatizada y de afectación significativa? Como vimos, existe un control *ex post* que se encarga de la eliminación de los falsos positivos. Es evidente que existe participación humana en el bucle, lo cual no obsta necesariamente que se excluya la aplicabilidad del precepto. Como se ha señalado, las directrices del GT 29 identifican 2 criterios relevantes para determinar si se cumple este requisito: en primer lugar, debe existir suficiente autoridad para alterar la decisión; asimismo, el humano debe tener acceso a todas las entradas pertinentes que dieron lugar a la clasificación[163]. En el caso, existe una autoridad suficiente, por lo que el primer requisito se cumple. Respecto del segundo, en cambio, es difícil saberlo, puesto que los procedimientos internos de cómo se seleccionan los casos en esta segunda etapa no son públicos.

Luego, si aceptamos que no es suficientemente relevante el grado de participación humana en el bucle, aún se debe evaluar si produce una afectación jurídica o similar. Ciertamente, el hecho de que una persona sea clasificada como sospechosa no parece alterar su condición jurídica, pues los efectos jurídicos serán posteriores a la investigación ulterior que se realiza. Empero, ser clasificado en situación de riesgo por un informe que se mantiene por dos años, y recibir a partir de él un mayor escrutinio de la autoridad, parece indicar que sí hay un efecto significativo en el sentido que apunta el RGPD. Aunque las propias directrices del GT 29 aclaran que es difícil determinarlo, señalan que un caso de afectación significativa sería la publicidad en línea, lo que da a entender que el estándar de "significativo" apunta a un universo bastante amplio[164]. Volveremos sobre este punto en análisis de la sentencia del Tribunal de la Haya.

---

en España tras el pronunciamiento del Tribunal Constitucional de España de 22 de mayo de 2019 (¿Legalizar las prácticas de Cambrigde Analytica?) (2020), p. 72. [Visible en: https://www.researchgate.net/publication/339376683_La_Elaboracion_de_Perfiles_Politicos_en_Espana_tras_la_Sentencia_del_Tribunal_Constitucional_de_22_de_Mayo_de_2019_por_la_que_se_Declara_Inconstitucional_el_Art_58_bis1_de_la_Ley_Organica_del_Regimen_].

160 Vid. GOODMAN, Bryce - FLAXMAN, Seth, *European Union regulations on algorithmic decision-making and a "right to explanation*", en *AI magazine* 38 (2017) 3, pp. 50-57.

161 "*En cualquier caso, dicho tratamiento debe estar sujeto a las garantías apropiadas, entre las que se deben incluir la información específica al interesado y el derecho a obtener intervención humana, a expresar su punto de vista, a recibir una explicación de la decisión tomada después de tal evaluación y a impugnar la decisión*." Vid. Considerando 71 del RGPD.

162 WACHTER *et al*, cit. (n. 76), pp. 79-82.

163 Grupo de trabajo sobre protección de datos del artículo 29, cit. (n. 152), p. 22.

164 Ibid., p. 24.

Continuando con el análisis del artículo 22, el apartado 4 señala que "*las decisiones a que se refiere el apartado 2 no se basarán en las categorías especiales de datos personales contempladas en el artículo 9, apartado 1, salvo que se aplique el artículo 9, apartado 2, letra a) o g), y se hayan tomado medidas adecuadas para salvaguardar los derechos y libertades y los intereses legítimos del interesado*". Esta norma se refiere a la toma de decisiones automatizada basada en el tratamiento de categorías de datos protegidas, (*i.e.* que revelen el "*origen étnico o racial, las opiniones políticas, las convicciones religiosas o filosóficas, o la afiliación sindical, y el tratamiento de datos genéticos, datos biométricos dirigidos a identificar de manera unívoca a una persona física, datos relativos a la salud o datos relativos a la vida sexual o las orientaciones sexuales de una persona física*")[165].

Notemos la relevancia para nuestro caso de estudio que exista una autorización, como no podía ser de otra forma, para utilizar estos datos de carácter sensible cuando *"el tratamiento es necesario para el cumplimiento de obligaciones y el ejercicio de derechos específicos del responsable del tratamiento o del interesado en el ámbito del Derecho laboral y de la seguridad y protección social* (…)"[166]. De este modo, pese a poder utilizarse excepcionalmente este tipo de datos en el tratamiento, la decisión automatizada del artículo 22 no se podrá basar en ellos.

b) Derechos de notificación y acceso

Aún si no existiese un derecho vinculante a la explicación derivado directamente del artículo 22, los artículos 13 a 15, leídos en su conjunto, parecen proporcionar una vía confortable para este propósito. Los derechos contenidos en los artículos 13 y 14 crean obligaciones de notificación; en el primer caso cuando se recojan datos personales directamente del individuo y en el segundo cuando se obtengan datos personales de otro lugar. El artículo 15, por su parte, crea un derecho del titular de acceder a la información relativa a sus datos. Exploremos esto con mayor detalle.

En lo que nos concierne, en virtud de los artículos 13, apartado 2, letra f), y 14, apartado 2, letra g) se establece la misma obligación de informar al interesado sobre "*la existencia de decisiones automatizas, incluida la elaboración de perfiles*" e "*información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado*". Esta disposición es aplicable "al menos" a los casos a los que se refiere el artículo 22, apartados 1 y 4. No obstante, como indican WACHTER *et al*, dado que este derecho parece referirse a la fase de recopilación de datos, la única información cronológicamente posible es la de un momento anterior al resultado de la clasificación del algoritmo[167]. En consecuencia, se trataría de un tipo de transparencia prospectiva, *i.e.* que busca informar sobre la "funcionalidad general del sistema" del algoritmo[168].

A diferencia de las disposiciones anteriores, el artículo 15 establece un derecho de acceso para los interesados. Este artículo establece que el interesado tiene un "*derecho de acceso a los datos personales y a la siguiente información*": (…) *h) la existencia de decisiones automatizadas, incluida la elaboración de perfiles, a que se refiere el artículo 22, apartados 1 y 4, y, al menos en tales casos, información significativa sobre la lógica aplicada, así como la importancia y las consecuencias previstas de dicho tratamiento para el interesado*"[169]. Como indican WACHTER *et al*., esta disposición es idéntica a la del artículo 13, apartado 2, letra f) y del artículo 14, apartado 2, letra g), por lo que podría pensarse que se refiere

[165] Artículo 9, apartado 1 del RGPD.
[166] Artículo 9, apartado 2, letra b).
[167] WACHTER *et al*, cit. (n. 76), pp. 82.83.
[168] Ibid.
[169] Artículo 15, apartado 2, letra f) del RGPD.

a un momento *ex ante*, pero a diferencia de dichos artículos, el derecho de acceso depende de la solicitud, por lo que puede efectuarse durante o después del procesamiento de datos[170]. En consecuencia, puede afirmarse que proporciona una protección a posteriori del tratamiento de datos.

Así, aunque descartemos la existencia del "derecho a la explicación" derivado del artículo 22, una lectura sistemática de los artículos 13 a 15 proporciona una base mucho más sólida para su construcción normativa. Además, no debe perderse de vista que a través de estas disposiciones sólo se está estableciendo un piso mínimo, y nada obsta de que se pueda ir aún más lejos en la información que se proporciona en función de la protección de los demás derechos. Queda por dilucidar, entonces, cuál es el contenido y alcance exigible de una explicación para que sea "significativa".

En este punto, las directrices del GT 29 son bastante sugerentes al señalar que "el responsable del tratamiento debe ofrecer al interesado información general (…) sobre los factores que se han tenido en cuenta para el proceso de toma de decisiones y sobre su peso respectivo a nivel global" de modo que le permita "impugnar la decisión"[171]. Es decir, esta disposición debe interpretarse de manera funcional. Así, con algún mayor grado de detalle, se ha planteado que el umbral de significación debe: i) "permitir a los interesados comprender la lógica seguida por el programa informático de toma de decisiones"; ii) "permitirles impugnar la decisión" (es decir, "verificar la legalidad y la veracidad de los datos utilizados, y comprobar sus efectos no discriminatorios"); y iii) "capacitar al destinatario de las decisiones para prever sus efectos"[172].

Sin embargo, también existen limitaciones relevantes respecto a su alcance. En primer lugar, el considerando 63 señala que, en relación con el derecho de acceso, "*no debe afectar negativamente a los derechos o libertades de otros, incluidos los secretos comerciales o la propiedad intelectual y, en particular, los derechos de autor que protegen el software*". Por ende, debe resguardarse el secreto comercial y la propiedad intelectual sobre el algoritmo. Sin embargo, a renglón seguido indica que "*estas consideraciones no deben tener como resultado la negativa a prestar toda la información al interesado*"[173].

A su vez, el GT 29 ha aclarado que los responsables del tratamiento no deben proporcionar "necesariamente una compleja explicación de los algoritmos utilizados o la revelación de todo el algoritmo"[174]. Del mismo modo, aclara que "la complejidad no es una excusa para no ofrecer información al interesado"[175]. Por lo tanto, la explicación "no debe ser tan detallada que resulte demasiado técnica para el interesado, ni demasiado intrusiva para afectar negativamente la protección del derecho a la propiedad intelectual"[176].

Por último, en el artículo 23 del RGPD se abordan otras posibles limitaciones adicionales a las obligaciones y derechos en virtud de los artículos 12 a 22. Estas limitaciones deben realizarse por la vía legislativa, respetando el contenido esencial de los derechos, y siendo necesarias y

---

170 Ibid.
171 Grupo de trabajo sobre protección de datos del artículo 29, cit. (n. 152), p. 30.
172 TABARRINI, Camilla, *Understanding the Big Mind. Does the GDPR Bridge the Human-Machine Intelligibility Gap?,* en *Forthcoming, EuCML – Journal of European Consumer and Market Law* 9 (2019) 4, p. 17. [Visible en: https://ssrn.com/abstract=3533225] [Consultado por última vez: 9/08/2021].
173 Considerando 63 del RGPD.
174 Grupo de trabajo sobre protección de datos del artículo 29, cit. (n. 152), p. 30.
175 Ibid., p. 28.
176 TABARRINI, cit. (n. 171), p. 19.

proporcionales en una sociedad democrática[177]. Entre estas causales, cabe destacar especialmente que el cumplimiento de los derechos pueda socavar "*objetivos importantes de interés público general de la Unión o de un Estado miembro, en particular un interés económico o financiero importante de la Unión o de un Estado miembro, inclusive en los ámbitos fiscal, presupuestario y monetario, la sanidad pública y la seguridad social*"[178].

c) Otras herramientas necesarias para la gobernanza algorítmica

Más allá de los derechos individuales, las medidas de transparencia dirigidas a propiciar el control de los grupos expertos constituyen la herramienta más crucial para resguardar la equidad del tratamiento automatizado de datos. En esta línea, cabe destacar especialmente la figura de la auditoría algorítmica. Como indica GOODMAN, "en ingeniería de seguridad, la auditoría identifica riesgos clave de los procesos, evalúa si se han establecido salvaguardas adecuadas y proporciona orientación sobre prevención de riesgos futuros"[179]. En esta línea, podemos distinguir "riesgos primarios (o inherentes)" y "secundarios"; los primeros se orientan a la "prevención de riesgos en el diseño inicial", mientras que los riesgos secundarios "buscan mitigar riesgos mediante medidas adicionales"[180]. Aplicando esto a la elaboración de perfiles, los enfoques preventivos pueden realizarse "antes, durante o después del procesamiento"[181].

Si bien es cierto que el RGPD no prevé explícitamente la realización de auditorías, la expresión de "medidas técnicas y organizativas" podría interpretarse como una orientación en este sentido. Así, por ejemplo, el artículo 24, apartado 1 establece la siguiente obligación: "*Teniendo en cuenta la naturaleza, el ámbito, el contexto y los fines del tratamiento, así como los riesgos de diversa probabilidad y gravedad para los derechos y libertades de las personas físicas, el responsable del tratamiento aplicará medidas técnicas y organizativas apropiadas a fin de garantizar y poder demostrar que el tratamiento es conforme con el presente Reglamento. Dichas medidas se revisarán y actualizarán cuando sea necesario*".

Además, existen otros instrumentos útiles que proporcionan un control experto en esta dirección. Entre estos, cabe destacar las EIPD del artículo 35 del RGPD[182]. En virtud de esta disposición, se requerirá realizar obligatoriamente antes del tratamiento de datos una EIPD, cuando sea probable que dicho tratamiento "*entrañe un alto riesgo para los derechos y libertades de las personas físicas*" (artículo 35, apartado 1)[183]. Esta se requiere específicamente (aunque no de forma exclusiva) en el caso de la elaboración automatizada de perfiles, cuando dicho perfil produce un efecto jurídico o similar (apartado 3, letra a) o se base en datos sensibles (apartado 3, letra b)[184].

---

177 Artículo 23, apartado 1.

178 Artículo 23, apartado 1, letra h) del RGPD. Cabe señalar además que se prevén otras excepciones en virtud del artículo 89, apartado 2, según el cual también pueden establecerse limitaciones a los derechos y obligaciones de procesamiento con fines estadísticos o de investigación científica o histórica. En la misma línea, el art. 89, apartado 3, aborda las limitaciones para el procesamiento con fines de archivo o en interés público.

179 GOODMAN, cit. (n. 42), p. 5-6.

180 Ibid.

181 Ibid.

182 Suelen indicarse también los sellos de privacidad, planes de certificación y códigos de conducta.

183 "*Cuando sea probable que un tipo de tratamiento, en particular si utiliza nuevas tecnologías, por su naturaleza, alcance, contexto o fines, entrañe un alto riesgo para los derechos y libertades de las personas físicas, el responsable del tratamiento realizará, antes del tratamiento, una evaluación del impacto de las operaciones de tratamiento en la protección de datos personales. Una única evaluación podrá abordar una serie de operaciones de tratamiento similares que entrañen altos riesgos similares."* Art. 35, apartado 1 del RGPD.

184 "*a) evaluación sistemática y exhaustiva de aspectos personales de personas físicas que se base en un tratamiento automatizado, como la elaboración de perfiles, y sobre cuya base se tomen decisiones que produzcan efectos jurídicos para las personas físicas o que les afecten*

Asimismo, según las directrices sobre decisiones automatizadas del GT 29 "una EIPD puede ser especialmente útil para aquellos responsables del tratamiento que no estén seguros de si sus actividades propuestas se ajustan a la definición del artículo 22, apartado 1, y, en caso de que una excepción identificada las permita, de qué medidas de protección deben aplicarse"[185]. De este modo, por ejemplo, en virtud de una EIPD se puede determinar si existe participación humana significativa en el bucle algorítmico y qué grado de afectación produce la clasificación del algoritmo en los derechos de los sujetos de datos, determinando con anticipación a su implementación la concurrencia de salvaguardas adecuadas para mitigar sus efectos.

IV. APLICABILIDAD DEL PRINCIPIO DE TRANSPARENCIA

Hemos llegado al momento de la síntesis. Tras el trabajo descriptivo y analítico realizado en las partes anteriores, se analiza el problema relativo a la aplicabilidad del principio de transparencia frente a la opacidad, especialmente el caso de la opacidad estratégica necesaria para impedir el juego con el sistema. En primer lugar, se analiza la importancia del principio de transparencia en el contenido de la sentencia de febrero de 2020, donde el Tribunal de Distrito de La Haya, basándose en los principios de la protección de datos contenidos en el RGPD, declaró que el uso de SyRI y la normativa que sirve de base para su despliegue (particularmente el artículo 65 de la Ley SUWI y el capítulo 5 letra a) del Decreto SUWI) es incompatible con el artículo 8, apartado 2 del CEDH. Seguido de ello, se ofrecen algunas observaciones basadas en el peso cualitativo del principio de transparencia en el caso. Y para finalizar, se discute acerca de las diversas posturas que se han sostenido en literatura informática y jurídica sobre transparencia y protección de datos, concluyendo con algunas ideas concretas y viables relativas a la aplicabilidad del principio de transparencia algorítmica.

1. Análisis del caso NJCM et al contra los Países Bajos

En el origen del asunto se encuentra una demanda dirigida contra el Estado de los Países Bajos, por parte de una coalición de organizaciones neerlandesas de derechos civiles. Los demandantes apelan al derecho a la protección de la vida privada, a la protección de la intimidad y a la protección de los datos personales, tal como se establece en el artículo 8 del CEDH, en los artículos 7 y 8 de la Carta, en el artículo 17 del Pacto Internacional de Derechos Civiles y Políticos ("PIDCP") y en la aplicación de esos derechos en la legislación comunitaria y nacional, especialmente en el RGPD[186]. Además, los denunciantes apelan al derecho a un "recurso efectivo" y al derecho a un juicio justo, como se estipula en los artículos 6 y 13 del CEDH, el artículo 47 de la Carta y el artículo 14 del PIDCP[187].

---

significativamente de modo similar." Artículo 35, apartado 2, a) del RGPD; "*b) tratamiento a gran escala de las categorías especiales de datos a que se refiere el artículo 9, apartado 1, o de los datos personales relativos a condenas e infracciones penales a que se refiere el artículo 10."* Artículo 35, apartado 2, b).

185 Grupo de trabajo sobre protección de datos del artículo 29, cit. (n. 152), p. 22.

186 Cit. (n. 3), párrafo 5 y ss. de la sentencia.

187 Ibid. También tomó participación como *amis curiae* el relator especial de la ONU sobre pobreza extrema y derechos humanos, Philip Alston, manifestando su preocupación por la violación a los derechos humanos en el uso de SyRI. El informe puede consultarse en: https://www.ohchr.org/Documents/Issues/Poverty/Amicusfinalversionsigned.pdf [Consultado por última vez: 9/08/2021]. La audiencia en el tribunal de distrito tuvo lugar el 29 de octubre de 2019 y la sentencia se pronunció el 5 de febrero de 2020.

El tribunal, siguiendo la tradición de su jurisprudencia, establece el marco de evaluación, examinando sucesivamente la protección de derechos humanos del CEDH, la protección de la legislación de la UE, la Carta y el RGPD, y finalmente, la relación mutua entre el CEDH y la legislación de la UE[188]. De este modo, el tribunal procede a examinar si la ley SUWI infringe el art. 8 del CEDH. En este punto, como indica HUESO, es preciso subrayar unos puntos claves del marco jurídico de los Países Bajos para entender el fallo: que la Constitución de los Países Bajos ("*Grondwet*") en su artículo 94 establece que el conflicto con las visiones favorables de los tratados que son vinculantes implica que las normas nacionales sean inaplicables[189]; que no tiene un control de constitucionalidad de las leyes encargada a un órgano específico (art. 120); y que su control se articula con relación al respeto del CEDH[190]. Por ende, como indica el autor, no se trata de un desconocimiento del RGPD por parte del Tribunal, sino que era la única vía para declarar inaplicable la ley a la luz del derecho superior[191].

Volviendo al fallo del tribunal, para determinar si la normativa SyRI infringe o no el derecho a la privacidad, el Tribunal abordó una concepción amplia de dicho derecho, determinando que está íntimamente relacionado con el derecho a la protección de los datos personales[192]. Así, bajo la premisa de que el CEDH no puede contener un alcance menor que el previsto en la Carta ni en el RGPD[193], la Corte determina que el RGPD es directamente aplicable, e interpretará el CEDH sobre la base de los principios generales contenidos en la regulación de protección de datos prevista en la Carta y en el RGPD[194].

Seguido de ello, el tribunal analiza la grado y severidad de la injerencia, para lo cual comienza por determinar qué es exactamente SyRI, y luego establecer si la elaboración de un informe de riesgo constituye un perfil y una toma de decisiones basada en un tratamiento únicamente automatizado, en el sentido que apunta el art. 22, apartado 1 del RGPD[195].

El tribunal contrasta las posturas de los intervinientes: por un lado, NJCM et al, basándose en un informe que la División Asesora del Consejo de Estado presentó al gabinete[196], sostienen que se trata de un proceso de "seguimiento digital a partir del cual se clasifica a los ciudadanos según un perfil de riesgo, haciendo uso de DL y minería de datos, a través de un enlace automatizado a gran escala, no estructurado y no dirigido, de archivos de grandes grupos de ciudadanos, procesando en secreto datos personales[197]; por otro lado, el Estado argumenta que cuando se usa SyRI, "(sólo) los datos de conjuntos de datos existentes de agencias designadas (gubernamentales) se comparan para descubrir discrepancias con el fin de verificar las afirmaciones del interesado, no mediante técnicas de DL, sino de un árbol de decisiones"[198].

---

188 Cit. (n. 3), párrafo 6.19 de la sentencia.

189 Art. 94 "*Los preceptos legales en vigor dentro del Reino no serán de aplicación, si la aplicación de los mismos fuere incompatible con estipulaciones de tratados y de acuerdos de organizaciones internacionales de derecho público que obligan a toda persona*". Constitución del Reino de los Países Bajos. [Visible en: http://roble.pntic.mec.es/jmonte2/ue25/holanda/holanda.pdf] [Consultado por última vez: 9/08/2021].

190 HUESO, Cotino, "*SyRI, ¿a quién sanciono?" Garantías frente al uso de inteligencia artificial y decisiones automatizadas en el sector público y la sentencia holandesa de febrero de 2020,* en *La Ley Privacidad* 4 (2020), apartado III.

191 Ibid.

192 Cit. (n.3) párrafos 6.23 y ss. de la sentencia.

193 Art. 52 de la Carta de la UE.

194 Cit. (n. 3) Párrafo 6.41 de la sentencia.

195 Ibid., párrafo 6.44 de la sentencia.

196 Ibid.

197 Ibid., párrafo 6.46 de la sentencia.

198 Ibid., párrafo 6.47 de la sentencia.

Pese a todo, ante la ausencia de información verificable, el tribunal llegó a la conclusión de que no puede determinar qué es SyRI. Después de todo, no se han hecho públicos los modelos ni indicadores de riesgo, ni se proporciona información adicional al tribunal al respecto[199]. Sin embargo, considera que "la legislación de SyRI también deja abierta la posibilidad de que se utilicen análisis predictivos, DL y minería de datos" al implementar SyRI, por lo que opina que el funcionamiento de SyRI al menos "encaja" con los sistemas de DL"[200]. No obstante, eso no constituye una prueba para el tribunal de que efectivamente se use aprendizaje profundo en el despliegue de SyRI[201]. Además, contrario a lo que expone la parte demandante, el tribunal establece que la normativa no implica la recopilación de datos no estructurados, puesto que, pese a ser extremadamente amplia, la legislación enumera exhaustivamente una lista de 17 categorías de datos[202]. Asimismo, la corte determina que no es relevante discutir acerca de si se trata de un proceso de *Big Data*, puesto que no habría una única definición, y aunque puede que sea el caso, de ella no se sigue ninguna consecuencia normativa relevante[203].

Luego, el tribunal repara en la opacidad del sistema respecto del afectado por la clasificación. En concreto, señala que "la legislación de SyRI no prevé la obligación de informar sobre esto a aquellos cuyos datos se procesan en SyRI", de modo que no se puede "esperar razonablemente que los interesados sepan que sus datos han sido procesados"[204]. Tampoco existe la obligación legal de poner proactivamente en conocimiento a interesados sobre el hecho de que se ha realizado un informe de riesgo sobre ellos[205].

Seguido de ello, el tribunal procede a determinar si existe una elaboración automatizada de perfiles y una decisión basada en el tratamiento únicamente automatizado que produce efectos jurídicos o una afectación significativa similar, en el sentido del apartado 1 del art. 22 del RGPD[206]. No hay que perder de vista que esta observación está orientada no a verificar si se cumplen las condiciones de legalidad establecidas en el RGPD, sino a determinar cuál es el nivel de injerencia que se produce en el derecho a la privacidad (en sentido amplio, como se ha expuesto). El tribunal estima que un informe de riesgo no puede considerarse que produce un efecto legal, pero sí una decisión individual automatizada que produce una afectación significativa similar; un informe de riesgo que se puede almacenar durante un período de 2 años no es, en este sentido, inocuo[207].

A partir de estas precisiones, el tribunal analiza si la injerencia puede encontrarse justificada de acuerdo con el artículo 8 inciso 2 del CEDH[208]. Como hemos viste anteriormente, los requisitos son que la injerencia esté prevista por ley, que sea proporcional y necesaria en una sociedad democrática.

En cuanto a que la aplicación de SyRI debe estar prevista por la ley, el tribunal determina (de acuerdo con la jurisprudencia del TEDH) que no se requiere una ley en sentido formal, sino que

199 Ibid., párrafo 6.49 de la sentencia.
200 Ibid., párrafo 6.51 de la sentencia.
201 Ibid.
202 Ibid., párrafo 6.50 de la sentencia.
203 Ibid., párrafo 6.52 de la sentencia.
204 Ibid., párrafo 6.54 de la sentencia.
205 Ibid.
206 Ibid., párrafo 6.55 y ss. de la sentencia.
207 Ibid., párrafo 6.59 de la sentencia.
208 Ibid., párrafos 6.66 y ss.

debe existir al menos una base en el derecho interno[209]. Asimismo, dicha base debe ser suficientemente accesible y previsible, de modo que el afectado pueda adaptar su comportamiento[210], lo que se corresponde con el examen de la "calidad de la ley"[211]. El tribunal toma como referencia el caso "S. y Marper contra el Reino Unido"[212][213] para determinar que los requisitos de accesibilidad y previsibilidad suponen que la normativa debe "proporcionar suficiente protección contra la arbitrariedad y con suficiente claridad la discrecionalidad"[214]. Por ende, las garantías que se proporcionan contra la injerencia deben estar previstas en la ley, y además deben ser suficientes para evitar abusos[215].

El tribunal establece que el requisito de calidad de la ley se encuentra estrechamente vinculado con un presupuesto más amplio, *id est* si es necesario en una sociedad democrática. En consecuencia, procede inmediatamente a verificar dicha evaluación.

En su examen, el tribunal —ciñéndose al margen de apreciación nacional que concede la jurisprudencia del TEDH— considera que el uso de SyRI se corresponde con una necesidad social imperiosa[216]. Lo anterior se deriva de la cuantificación del fraude anual a la seguridad y asistencia social y los fines que persigue su uso, vale decir, el bienestar nacional económico. Por lo tanto, no se puede afirmar a priori que sea desproporcionado en aras del fin que persigue.

A continuación, el tribunal valora si se cumple con la "necesidad, proporcionalidad y subsidiariedad" requeridos por el artículo 8, inciso 2, del CEDH. En este punto, para la determinación de un equilibrio justo, el tribunal analizará el cumplimiento de los principios de la protección de datos[217]. El tribunal tiene en cuenta la gran cantidad de datos elegibles para el procesamiento de SyRI, la opacidad de los modelos de riesgo, indicadores y datos utilizados, el margen para ajustar el modelo en función de los resultados de la retroalimentación y, por último, el desconocimiento de los afectados sobre la existencia de un informe de riesgo, cuyo efecto se considera "significativo"[218].

Aquí el tribunal hace suya una consideración del mencionado caso S. y Marper contra el Reino Unido, donde se establece "que cualquier Estado que reclame un papel pionero en el desarrollo de nuevas tecnologías tiene la responsabilidad especial de lograr el equilibrio adecuado en este sentido", pues si bien considera que el legislador neerlandés no sería pionero respecto de SyRI, la creciente importancia del derecho a la protección de datos y el análisis de datos utilizando las nuevas tecnologías también sujetan al legislador a una responsabilidad especial[219].

[209] Ibid., párrafo 6.66 de la sentencia.
[210] Ibid.
[211] Este mismo principio de "calidad de la ley" también se aplicó en el caso STJUE Digital Rights C-293/12 y C-594/12 (FJ.65).
[212] El caso S. y Marper contra el Reino Unido se refería a la legalidad de la Ley de Protección de Datos del Reino Unido (1998), que implementó la Directiva 95/45 y las directrices basadas en ella para el uso de la Computadora Nacional de la Policía. Vid. S. and Marper v. The United Kingdom [GC], n. º 30562/04 y 30566/04, 04.12.2008.
[213] Cit. (n.3.) párrafo 6.68. de la sentencia.
[214] Ibid., párrafo 6.69 de la sentencia.
[215] Ibid., párrafo 6.70 de la sentencia.
[216] Ibid., párrafo 6.76 de la sentencia.
[217] Ibid., párrafo 6.82 de la sentencia.
[218] Ibid.
[219] Ibid., párrafo 6.84 de la sentencia.

El tribunal establece que el "principio de transparencia es el principio rector de la protección de datos"[220]. Así, dado que no se proporciona información alguna sobre los datos que pueden determinar la presencia de una circunstancia que amerite mayor riesgo, el tribunal opina que no se ha respetado suficientemente en la legislación este principio[221]. Además, la normativa no proporciona información sobre el funcionamiento del modelo de riesgo, y más específicamente, sobre el tipo de algoritmos utilizados, ni el método de análisis por parte de la SZW[222]. Tampoco se informa sobre los procesos de validación del modelo, ni la verificación de los indicadores de riesgo[223]. Todo ello se traduce en que no es posible comprobar cómo surge y de qué etapas se compone el árbol de decisiones referido por el Estado, por lo que determina que es difícil (si no imposible) imaginar cómo el afectado por la clasificación puede defenderse de la elaboración de un informe de riesgo, ni saber si sus datos se han procesado de forma correcta[224]. Además, considera que el hecho de destruirse los datos que no den lugar a un informe de riesgo en un período de tiempo no compensa la falta de transparencia[225].

El tribunal también destaca la transparencia en aras de la verificabilidad, puesto que es posible que el uso del modelo y análisis impliquen efectos discriminatorios involuntarios, especialmente en los análisis de datos basados en la elaboración de perfiles[226]. También se indica el efecto estigmatizador alegado por el Relator Especial de Naciones Unidas, dado el hecho de que SyRI suele emplearse en los vecindarios más pobres —áreas presuntamente problemáticas— lo que puede traducirse en un efecto de retroalimentación negativa, en la medida en que existirá una mayor probabilidad de encontrar una mayor cantidad de irregularidades en dichos lugares[227].

En esta línea, si bien no considera que el empleo de SyRI en áreas problemáticas sea en sí mismo desproporcionado, dada la gran cantidad de datos susceptibles de ser empleados en la elaboración de perfiles, el tribunal reconoce que existe el riesgo real de que SyRI opere con alguna clase de sesgo, identificando *proxys* entre el estatus socioeconómico o el antecedente de inmigración como un factor que indique mayor probabilidad de riesgo[228]. La normativa tampoco permite valorar si este riesgo se aborda suficientemente[229].

Igualmente, aunque se prevé un control humano para falsos positivos y negativos, el tribunal considera que es insuficiente, pues no se informa cómo se seleccionan los casos probables[230]. Por lo demás, la legislación sólo considera una supervisión retrospectiva de la autoridad de protección de datos.

A la luz de estas precisiones, el tribunal determina que no se han previsto suficientes garantías para proteger el derecho al respeto a la vida privada; sin comprensión de los indicadores y modelos, ni otras salvaguardas compensatorias, no puede afirmarse que la normativa proporciona suficiente protección contra la interferencia en este derecho[231].

---

220 Ibid., párrafo 6.87 de la sentencia.
221 Ibid.
222 Ibid., párrafo 6.89 de la sentencia.
223 Ibid.
224 Ibid., párrafo 6.90 de la sentencia.
225 Ibid.
226 Ibid., párrafo 6.91 de la sentencia.
227 Ibid., párrafo 6.92 de la sentencia.
228 Ibid., párrafo 6.93 de la sentencia.
229 Ibid., párrafo 6.94 de la sentencia.
230 Ibid.
231 Ibid., párrafo 6.95 de la sentencia.

El tribunal, además, considera que no se han respetado suficientemente los principios de limitación de finalidad y minimización de datos[232]. Si bien existe una delimitación clara de la finalidad perseguida, la gran cantidad de datos disponibles, sumado a la ausencia de una prueba de necesidad realizada por un tercero independiente, y no sólo de las autoridades designadas como prevé la normativa impugnada, produce una afectación a dichos principios[233]. Además, la lista que establece la normativa, si bien es exhaustiva en su enumeración, es tan amplia que sería "casi imposible concebir datos personales que no sean elegibles para el procesamiento de SyRI"[234]. De forma análoga, considera que la prueba de necesidad prevista en la normativa solo se efectúa por cada organismo respecto de sus propias bases de datos, sin considerarse una prueba integrada de antemano, idealmente realizada por un tercero[235].

En la misma línea, tiene en cuenta la opacidad de los modelos e indicadores de riesgo, pues sin información verificable es imposible valorar si el suministro de datos es necesario, y en qué medida[236]. En consecuencia, concluye que "el interesado no tiene la certeza suficiente de que su privacidad está garantizada cuando se utiliza SyRI"[237].

Un último aspecto importante que se discute es la realización de una EIPD, en el sentido previsto por el artículo 35, apartado 1 del RGPD. El tribunal considera, junto al Estado, que esta disposición no es directamente aplicable, toda vez que se ha llevado a cabo dicha evaluación el contexto de la tramitación de la ley[238]. Sin embargo, de ello no se sigue que no deba realizarse dicha evaluación antes de la realización de cada proyecto SyRI[239]. Además, la EIPD realizada se llevó a cabo antes de la entrada en vigor del RGPD. Por tanto, si bien no puede juzgar sobre la base de la información disponible si dicha evaluación cumple con el requisito previsto en el art. 35 del RGPD, es claro que la ausencia de ésta en cada proyecto agrava la injerencia que se produce en el derecho al respeto a la privacidad del artículo 8, inciso 2, del CEDH[240].

En vista de todo lo anterior, el Tribunal consideró incompatibles con la sentencia el artículo 65 de la Ley SUWI y el capítulo 5, letra a, del Decreto SUWI, por contravenir el artículo 8, apartado 2, del CEDH a la luz los principios fundamentales en los que se basa la protección de datos consagrados en el Derecho de la Unión, y especialmente el principio de transparencia.

Frente a la falta de precedentes en la materia, esta sentencia resulta de especial interés para el análisis del principio de transparencia y su aplicabilidad, pues pese a pronunciarse directamente sobre el derecho a la privacidad, también plantea un paso decisivo en la clarificación de la relación entre el derecho a la protección de datos y el principio de transparencia. De este modo, con este fallo como telón de fondo, es posible extrapolar estas conclusiones a la lógica exclusiva de la elaboración de perfiles, el derecho a la protección de datos y la transparencia. A continuación, se abordan los aspectos que merecen mayor atención a partir del estándar de transparencia que plantea el tribunal en su sentencia en función de nuestro objeto de estudio.

[232] Ibid., párrafo 6.96 de la sentencia.
[233] Ibid., párrafo 6.97 de la sentencia.
[234] Ibid., párrafo 6.98 de la sentencia.
[235] Ibid., párrafo 6.99 de la sentencia.
[236] Ibid., párrafo 6.100 de la sentencia.
[237] Ibid.
[238] Ibid., párrafo 6.104 de la sentencia.
[239] Ibid., párrafo 6.105 de la sentencia.
[240] Ibid.

Es importante destacar que el tribunal considera que el uso de herramientas como SyRI no es, en principio, jurídicamente problemática. Es más, su postura fue favorable, ya que vio en ella una tarea que coincide con una necesidad social acuciante. Contrario a otras lecturas de esta sentencia, parece ser que el estándar de valoración que otorga el tribunal concuerda con la visión que plantea HUESO, en el sentido de que sería no obstaculizadora al despliegue de esta tecnología, pero sí condicionada a un "alto estándar de garantías"[241].

Dicho estándar, ciertamente, encuentra su base en que los Estados tendrían una "responsabilidad especial cuando aplican nuevas tecnologías"[242]. Así las cosas, la sentencia no hace más que corroborar, en su extensión, el carácter rector del principio de transparencia en el derecho a la protección de datos, como se expresa en el fallo[243]. De este modo, el tribunal centra el mayor peso de su argumentación en la opacidad intencional que rodea todo el sistema. Por consiguiente, pese a los resguardos tomados a la hora de implementar SyRI, teniendo presente las numerosas garantías que se ofrecían, la opacidad del sistema es tal que termina por hacer decaer la legitimidad del sistema. Esta noción general la podemos desglosar en varios razonamientos subyacentes a partir de las barreras de opacidad analizadas anteriormente.

En primer lugar, visto desde el punto de vista del afectado por la clasificación de riesgo, el fallo se esfuerza en recalcar la importancia de la notificación "puerta a puerta", pues de lo contrario sería imposible impugnar la clasificación de riesgo. La opacidad intencional del sistema plantea, por ende, importantes barreras para que el afectado pueda evaluar su exactitud, equidad, legalidad y la consiguiente oportunidad de impugnación. De este modo, se manifiesta insuficiente la transparencia pasiva relativa a informar el mero resultado por la clasificación. Se requeriría una notificación (*i.e.* informar de manera proactiva) y algún tipo de retroalimentación sobre cómo se llegó a dicho resultado. No está demás subrayar que lo anterior supone, a su vez, superar el umbral de la opacidad alfabética para que la información sea inteligible.

Este planteamiento concuerda con el requisito que introducen los artículos 13. 2º f), 14. 2º g) y 15 h) del RGPD al añadir la obligación de proporcionar información significativa sobre la lógica aplicada y las consecuencias previstas del tratamiento de datos. A esto se debe sumar el formato establecido por el artículo 12 (*i.e. en forma concisa, transparente, inteligible y de fácil acceso, con un lenguaje claro y sencillo*).

Aunque no reparó mucho en ello, la sentencia parece dar cuenta de que el tratamiento que realizaba SyRI era una decisión "únicamente automatizada", en el sentido del artículo 22 del RGPD. Lo anterior, pues la eliminación de los falsos positivos no parecía entenderse como una "intervención humana significativa". Además, el tribunal consideró que la afectación que deriva de un perfil de riesgo, si bien no producía efectos jurídicos, sí constituye una afectación significativa[244]. Pues pese a que no era la base normativa que se estaba aplicando, el tribunal hizo hincapié en este punto para subrayar que, a la luz de la gravedad de la injerencia que se produce, se requieren garantías especiales para compensar dicha afectación que deriva de una decisión que incluye la elaboración de perfiles únicamente automatizada con una afectación significativa.

---

241 HUESO, Lorenzo, *Hacia la transparencia 4.0, el uso de la inteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales,* en RAMIÓ, Carles (coord.), *Repensando la Administración digital y la innovación pública* (Instituto Nacional de Administración Pública INAP, Madrid, 2021), p. 10.

242 Cit. (n.3), párrafo 6.84 de la sentencia.

243 Ibid., párrafo 6.87 de la sentencia.

244 Ibid., párrafo 6.57 de la sentencia.

De este modo, bajo la óptica de responsabilidad proactiva del RGPD, cuando se trata de decisiones automatizadas del artículo 22 RGPD, se refleja un compromiso normativo firmemente garantista con el sujeto de datos, por lo que las medidas compensatorias deberían ser mucho mayores cuando se trata de decisiones de alto riesgo. Sin embargo, esta visión garantista que aquí se sigue, como se ha expuesto, no es absoluta, y se contrapone a las posibles limitaciones autorizadas por el art. 23 del RGPD en aras del interés público. Tales restricciones, huelga decir, son de aplicación estricta, y proceden solamente en la medida en que sean "proporcionales y necesarias en una sociedad democrática" y se proporcionen garantías suficientes para proteger los demás derechos. En consecuencia, se requiere una adecuada evaluación a la luz de todos los valores en competencia para determinar si es válida la limitación establecida.

De este modo, tal como señaló el tribunal, en el caso en concreto de SyRI no había proporcionalidad entre la injerencia y las garantías proporcionadas. Es decir, la opacidad frente a los sujetos de datos fue relevante para determinar que el uso y la normativa SyRI no cumplen con los estándares en materia de derechos humanos.

Una segunda advertencia queda de manifiesto sobre la opacidad intencional y el control de terceros expertos. Esta preocupación se concreta principalmente por la falta de auditorías externas e independientes, como también de la realización de las EIPD de manera periódica.

En el caso de estas últimas, si bien no constituía aparentemente una infracción al RGPD su no realización periódica, se plantea que, de haberse llevado a cabo antes de cada proyecto SyRI, podría haberse mitigado (en parte) el elevado grado de injerencia que se producía en los derechos de los afectados. Asimismo, el tribunal tuvo presente que el uso de SyRI puede estar operando con sesgos y, por ende, que tuviera efectos discriminatorios sobre minorías históricamente desfavorecidas, especialmente respecto de atributos como la raza, la etnia y el estatus socioeconómico, dada la lógica de "barrio" (*i.e.* basada en el código postal) con que se empleaba SyRI. En este sentido, las auditorías algorítmicas constituyen una garantía fundamental de los derechos del afectado, en aras de la comprobación del funcionamiento de los algoritmos utilizados para el tratamiento automatizado de sus datos personales. Por ende, la falta de transparencia dirigida al control experto también es relevante para determinar una afectación ilegítima a los derechos fundamentales de los afectados.

Un tercer aspecto importante para destacar es la cuestión relativa a la interpretabilidad del modelo. Como se ha expuesto anteriormente, la opacidad puede ser no intencional, en el caso de la opacidad alfabética e intrínseca. En cuanto a los algoritmos empleados en el despliegue de SyRI, la opacidad deliberada hizo imposible al tribunal verificar el funcionamiento interno y la familia de los algoritmos. Sin embargo, el tribunal afirmó que el uso de SyRI encajaba con sistemas de DL y que, por lo demás, la ley que autorizaba su uso dejaba la puerta abierta para que ello sucediera.

Es un debate abierto hasta qué punto la lucha por derribar eficazmente los límites de opacidad requiere de interpretabilidad de los algoritmos que se emplean. Podría plantearse, sin embargo, que en decisiones de alto riesgo como SyRI, si la preocupación por los derechos fundamentales exige un mayor control o previsibilidad *ex ante*, su cumplimiento no podría quedar supeditada a la viabilidad técnica; si resulta que la protección de los derechos exige algo que cierto tipo de algoritmos no pueden proporcionar, entonces no pueden utilizarse sin vulnerar tales derechos. Sin embargo, ante la falta de otras interpretaciones jurisprudenciales, sería infructuoso urdir más en ello.

En síntesis, se pueden elucubrar una serie de razones para pedir transparencia en el ciclo algorítmico de la elaboración automatizada de perfiles. Pocas dudas caben de que garantizar el derecho a la protección de datos de los ciudadanos requiere que exista suficiente transparencia para que el tratamiento de datos a través del sistema pueda ser enjuiciado por los sujetos de datos o terceros expertos en caso de presentar indicios de error, sesgo o de discriminación (directa e indirecta). De igual forma, la Administración tiene el deber de establecer garantías compensatorias suficientes para mitigar estos riesgos.

Por consiguiente, la falta de auditorías y EIPD, la opacidad respecto a los afectados directamente por la clasificación y, en general, la opacidad sobre el modelo y los indicadores de riesgo —lo que podría incluir eventualmente un escrutinio sobre la opacidad intrínseca del modelo—, constituyen razones más que evidentes para determinar que la opacidad del ciclo algorítmico, cuando se perfilan personas físicas y existiendo una afectación significativa como resultado de ese proceso, constituyen elementos relevantes para considerar que se produce una afectación ilegítima a los derechos de los titulares de datos. En otras palabras, la opacidad algorítmica puede ser determinante para verificar que existe una vulneración al derecho a la protección de los datos personales de los sujetos de datos.

Esta conclusión preliminar, sin embargo, nos conduce al problema del quantum de la transparencia, vale decir, en qué medida la opacidad del ciclo algorítmico puede ser tolerada sin afectar desproporcionadamente el derecho a la protección de datos. Queda por evaluar, por tanto, hasta qué punto la aplicación del principio de transparencia puede derrotar las pretensiones que sustentan la opacidad intencional. Las siguientes consideraciones tendrán como objetivo identificar un método adecuado para ponderar estos valores.

2. Entre Escila y Caribdis: transparencia vs. opacidad estratégica

Vimos hasta ahora que la transparencia encierra promesas difíciles de realizar en aplicaciones concretas. Evidenciamos, asimismo, una clara tensión entre dos valores en competencia: la transparencia y la opacidad necesaria para impedir el juego con el sistema. En este orden de cosas, cabe dilucidar todavía qué tipo de transparencia es posible, en qué situaciones y para quien puede ser procedente. En concreto, qué elementos constitutivos de cada fase del proceso (A, B, y C de la tabla 2) deben revelarse y respecto de qué grupo objetivo (identificados como 1, 2, y 3 en la tabla 2), dando cuenta, a su vez, qué tipo de vías positivas son plausibles para conseguir dicho objetivo. A *contrario sensu*, este examen nos permitirá, con algún mayor grado de detalle, vislumbrar en qué medida la opacidad deliberada puede ser tolerada sin constituir una afectación ilegítima al derecho a la protección de datos personales. Concluida esta sección, será posible avizorar los contornos de la aplicación efectiva del principio de transparencia para el caso.

En línea con lo señalado en secciones anteriores a esta, la transparencia puede resultar problemática en un sentido relevante —constituyendo un fundamento para la opacidad intencional— al menos en 3 sentidos: i) por la afectación a la ventaja competitiva y el secreto comercial; ii) por el resguardo a la privacidad de los afectados por la clasificación y la seguridad de sus datos personales; y iii) por la opacidad estratégica requerida para impedir el juego[245]. Todas

[245] DE LAAT menciona que suele argüirse como un contraargumento adicional el problema de la opacidad intrínseca de los algoritmos. Sin embargo, para efectos de este trabajo, pese a que suele encapsularse la complejidad del modelo como una justificación a favor de la opacidad deliberada, no se considera como un argumento propiamente tal, sino de una característica que cualquier política de transparencia algorítmica deberá considerar, como también ocurre con la dimensión alfabética de la opacidad. Por tanto, no se abordará como un elemento independiente, sino como

ellas son objeciones válidas, y ciertamente desalientan cualquier intento de una transparencia total accesible al público en general (*i.e.* fases A, B y C, dirigidas al grupo 1 de la tabla 2), que sería la respuesta intuitiva —aunque algo reaccionaria— que se podría urdir frente a la opacidad de los sistemas algorítmicos. Además, esta postura ha sido defendida en la literatura, sin éxito, ya en diversas ocasiones, por lo que sería infructuoso conjeturar más en ello[246].

Empero, también vimos que sería improcedente un grado de opacidad tal que impida: por un lado, la información relativa al tratamiento de datos respecto de los afectados, pues debería existir al menos una posibilidad en que puedan defenderse de una clasificación adversa; por otro, la no verificación de auditorías y la no realización de EIPD.

Así, la posición que se deriva desde la lógica de la protección de datos a partir del trabajo analítico desarrollado parece ser mucho menos ambiciosa que una "transparencia total" accesible al público general. Parece más razonable intentar alcanzar un balance óptimo del cumplimiento del principio de transparencia, al menos, en dos niveles; por un lado, residenciando la "transparencia total" —*i.e.* que incluya todas las fases del proceso— a los grupos expertos, manifestada a través de auditorías y EIPD, y, por otro, suministrando información significativa sobre la fase de implementación del proceso (C de la tabla) a los sujetos de datos afectados adversamente por la clasificación (designados como parte del grupo 3 de la tabla) —*i.e.* respecto de aquellos cuya clasificación (negativa) daría lugar a un informe de riesgo, por lo que habría una afectación similar a un efecto jurídico en el sentido del artículo 22 del RGPD—. Se trata de una posición intermedia que, sin caer en un enfoque prohibicionista, permite una articulación sociotécnica coherente con los derechos individuales de transparencia que se derivan del RGPD, el estándar de garantías establecidas en el fallo analizado y el resguardo de la funcionalidad del sistema de cara a la opacidad estratégica[247]. Por tanto, cabe hacer eco de las críticas que suelen plantearse a las políticas de transparencia, y con tales precisiones en mente, abordar los contraargumentos pertinentes para terminar de delimitar esta postura.

En primer lugar, como se ha visto, la regulación existente en materia de protección de datos exige no afectar negativamente los derechos de terceros, y en particular, la propiedad intelectual de los desarrolladores[248]. Desde esta perspectiva, especialmente en lo referido al código fuente, el cumplimiento de la transparencia superaría el estándar concedido por el RGPD. Una forma de abordar esta disyuntiva podría ser la que plantea BOIX, en el sentido de que el Estado debería

una característica perenne que deberá tomar cuenta cualquier noción de transparencia. Vid. DE LAAT, Paul, cit. (n. 29). En la misma línea, refutando la idea de la complejidad como un argumento para mantener en secreto el código fuente, vid. BOIX, Andrés, *Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones,* en *Revista de Derecho Público: Teoría y Método* 1 (2020) 33, p. 264.

[246] Un sofisticado intento por conseguirla puede consultarse en: ZARSKY, cit. (n. 75), pp. 1523 y ss. En la misma línea, vid. DE LAAT, cit. (n. 29).

[247] Excede el propósito de esta memoria buscar un equilibrio entre transparencia y el juego con el sistema más allá del contenido informacional que se pueda proporcionar a diversos grupos objetivos de las políticas de transparencia. No obstante, cabe enfatizar que existen otras medidas complementarias para resguardar la funcionalidad del sistema. En esta línea, a propósito de las oportunidades que ofrece el Big Data, podemos mencionar: agregar más variables proxy o introducir más aleatoriedad, cambiar frecuentemente el propio modelo de riesgo, aumentar el peso de indicadores inmutables, o bien utilizar más datos, con diferentes fuentes, sobre el mismo conjunto de sujetos. Vid. BAMBAUER - ZARSKY, cit. (n. 130), pp. 14 – 15.

[248] Considerando 63 del RGPD.

adquirir el programa y no la "mera licencia de uso", o bien creándolos por cuenta propia[249]. Sin embargo, como señala el propio BOIX, aunque dicho planteamiento pueda ser atendible, el estado del arte de la protección de datos personales no parece ser la vía para exigir este estándar, al menos a falta de interpretaciones jurisprudenciales en esa orientación[250].

Sin perjuicio de ello, al menos respecto de los grupos expertos, la información relativa al código siempre puede suministrarse resguardando el secreto comercial, con fines de auditoría o EIPD, a través de cláusulas de confidencialidad. Por parte de los sujetos de datos, en cambio, este argumento resulta absolutamente pertinente.

En la misma línea, como hemos venido reiterando, la transparencia, para considerarse significativa, debe considerar las circunstancias individuales del receptor, en aras de privilegiar la comprensibilidad del grupo objetivo, por lo que se requiere una explicación simple y, precisamente, de carácter no-técnico. Siendo esto así, aparece como absolutamente desproporcionada, la idea de revelar información cubierta por el secreto comercial a los sujetos de datos. Además, como ya mencionamos, el respeto por el secreto comercial no constituye una excusa suficiente para no cumplir las obligaciones en materia de transparencia. Parece indicado, entonces, restringir la divulgación completa del algoritmo, así como la información relevante sobre todas las fases del proceso, únicamente al control experto. Por ende, este argumento no es suficiente para desestimar la posición intermedia antes mencionada.

La segunda crítica refiere, basándose también en los derechos de terceros, que la transparencia sería problemática al tener en cuenta el derecho a la privacidad de los sujetos de datos. Ciertamente, la divulgación de los datos de entrada que se han procesado, así como también el resultado de la inferencia, podría tener un efecto estigmatizador y afectar negativamente este derecho[251]. Asimismo, como hemos mencionado, su publicidad también implicaría hipotecar la seguridad de los datos, puesto que terceros podrían almacenarlos y usarlos para sus propios fines, lo que lógicamente contradice el espíritu del derecho a la protección de datos[252].

Ahora bien, este argumento tampoco permite declinar nuestra propuesta, puesto que sólo es atingente a las políticas de transparencia orientadas al público en general (grupo 1 de la tabla 2). En el caso de los grupos expertos, cuando se realizan tareas de control o supervisión, las medidas de protección de datos deben considerar necesariamente técnicas de seudonimización y seguridad[253], por lo que no sería un aspecto problemático. En cuanto a los sujetos de datos, lógicamente no sería un problema, en la medida en que la información se restrinja a sus propios datos personales.

Debemos abordar ahora la objeción principal a nuestra postura, *id est* la cuestión referida a la opacidad estratégica. Como vimos, este argumento consiste en que la transparencia puede invitar a los interesados a jugar con el sistema, entendiendo como juego que el individuo, a través de un

---

[249] BOIX, Andrés, *Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones,* en *Revista de Derecho Público: Teoría y Método* 1 (2020) 33, p. 256.

[250] Ibid., p. 244-249.

[251] ZARSKY, cit. (n. 75), pp. 1560 y ss.

[252] DE LAAT, cit. (n. 29), pp. 535-536.

[253] Así lo dispone, por lo demás, el RGPD cuando señala: "1. *Teniendo en cuenta el estado de la técnica, los costes de aplicación, y la naturaleza, el alcance, el contexto y los fines del tratamiento, así como riesgos de probabilidad y gravedad variables para los derechos y libertades de las personas físicas, el responsable y el encargado del tratamiento aplicarán medidas técnicas y organizativas apropiadas para garantizar un nivel de seguridad adecuado al riesgo, que en su caso incluya, entre otros: a) la seudonimización y el cifrado de datos personales*". Vid. Artículo 32 del RGPD.

cambio de su conducta y sin alterar la característica clave que se intenta predecir, logrará a voluntad modificar la inferencia arrojada por el algoritmo[254]. De este modo, la transparencia afectaría la correcta funcionalidad del sistema. Este argumento es atendible únicamente respecto de los individuos susceptibles del juego, es decir, los sujetos de datos afectados por la clasificación, por lo que, en lo que al control experto se refiere, nuestra postura se mantiene indemne.

En la literatura más reciente, la tendencia ante esta disyuntiva ha sido declinar la balanza en perjuicio de la transparencia. Así, por ejemplo, BOIX señala —aunque refiriéndose a un presupuesto distinto, a saber, la publicidad el código fuente— que en tareas como la investigación del fraude esta observación "sí constituye un argumento crítico de más interés jurídico", para luego sugerir que debería primar la opacidad, tal como sería en un "entorno no electrónico"[255]. En la misma línea, TODOLÍ señala —a raíz del caso SyRI— que deberían desarrollarse "controles internos y externos" para resguardo de los derechos involucrados, pero que sólo debería ser exigible la publicidad sobre "qué datos son tenidos en cuenta, de forma general, por el *Big Data* para crear el índice de riesgo", con el fin de identificar el uso de información protegida o discriminatoria en un sentido jurídicamente problemático, pues con más información los individuos podrían jugar con el sistema[256]. No tiene mucho sentido seguir profundizando en ello. Basta con señalar que, en buena medida, las posiciones gravitan esencialmente en la siguiente idea: atendida la función que cumple el algoritmo (*i.e.* la prevención y detección del fraude), la opacidad estratégica debe primar por sobre los intereses individuales de los afectados adversamente por la clasificación en materia de transparencia. No así respecto de las medidas orientadas a propiciar el control experto, donde parece haber bastante consenso de su importancia.

A primera vista, se trata de un pleito no muy reñido, pues la opacidad en este caso no parece en modo alguno ser "desproporcional" para restringir los derechos individuales de transparencia, en atención a las posibilidades de limitación que establece el artículo 23 del RGPD. Las observaciones planteadas pueden compartirse. Sin embargo, hay un argumento sencillo que puede ser problemático: como indica el propio TODOLÍ, si asumimos que el estándar de transparencia es tan alto que en definitiva termina por frustrar cualquier intento de funcionalidad, entonces es válido interpretar que se estaría "prohibiendo de facto" el uso de la herramienta[257].

En esta línea, TODOLÍ argumenta que sería factible proporcionar información sobre el tipo de datos que se utiliza. Sin embargo, esto no es más de lo que ya se venía haciendo a través del decreto SUWI. Y vimos con algún grado de detalle, de la mano de la sentencia del Tribunal de la Haya, que este grado de opacidad sería manifiestamente "desproporcional" a la luz de los derechos de los afectados por la inferencia, pues se requeriría, además del control experto, algún tipo de retroalimentación dirigido a los sujetos de datos que les permita defenderse de la clasificación. Por ende, la solución alejandrina a este nudo gordiano —que sería la de asumir prima facie que no es posible proporcionar información significativa sobre la lógica interna del tratamiento de sus datos al afectado por la clasificación— podría traer como consecuencia lógica la prohibición de facto. No obstante, esta interpretación encierra una contradicción con la perspectiva jurisprudencial y la que aquí se ha defendido: que la elaboración de perfiles de riesgo

[254] BAMBAUER – ZARSKY, cit. (n. 130), pp. 6-11.
[255] BOIX, cit. (n. 248), pp. 264-265.
[256] TODOLÍ, Adrian, *Retos legales del uso del big data en la selección de sujetos a investigar por la Inspección de Trabajo y de la Seguridad Social,* en *Revista Galega de Administración Pública* 59 (2020) 1, p. 334.
[257] Ibid., p. 329.

es una empresa perfectamente legítima y, más aún, en palabras del propio tribunal, de una "necesidad social imperiosa"[258].

Una segunda lectura plausible sería la de asumir que bastaría para satisfacer el principio de transparencia el control experto a través de auditorías y EIPD. Sin embargo, esta interpretación obtura el costo que ello podría significar, pues como indicó el tribunal, este nivel de opacidad respecto de los afectados aún puede tener un gran impacto negativo en la confianza de los ciudadanos, lo que podría provocar un "efecto paralizador" en la disposición de compartir sus datos personales[259]. Después de todo, como se señaló anteriormente, los datos son esenciales para que las nuevas tecnologías sean funcionales, por lo que la confianza del ciudadano es también fundamental. En consecuencia, dado que la transparencia también se orienta a reforzar la confianza —esencial para la funcionalidad de este tipo de sistemas—, puede que el grado de opacidad expuesto también termine mermando la funcionalidad del sistema.

Empero, es posible discrepar de dichas interpretaciones y proporcionar una tercera lectura: sería posible proporcionar información sobre la lógica aplicada en el tratamiento de datos, desde la perspectiva de la transparencia retrospectiva (*i.e. ex posteriori* a una decisión específica), que proporcione algún efecto indiciario en favor de los sujetos de datos adversamente afectados por la clasificación, siempre en el entendido de que el contenido informacional debe ser reducido al mínimo para que no afecte negativamente la funcionalidad del sistema. Entonces sería el caso que se satisface, por un lado, el interés que dimana de los derechos individuales de transparencia y, por otro, el requisito de la funcionalidad. Se trata, pues, de una medida complementaria, dado que la transparencia total residiría únicamente en el control experto. De este modo, el punto de equilibrio debería buscarse en una fórmula que reconcilie el distanciamiento entre este estándar de garantías en materia de transparencia y la funcionalidad del sistema, y no en descartar tan a la ligera como algunos autores han planteado.

En el corazón del desafío por armonizar estos dos valores en tensión subyace, no obstante, una doble dificultad: por un lado, que la explicación tenga un contenido informacional mínimamente gravoso, con el fin de evitar que los sujetos de datos puedan manipular el sistema; por otro lado, la explicación debe resultar significativa, tomando en consideración que debe ser inteligible a la luz de las barreras de opacidad alfabética e intrínseca.

En cuanto al primer criterio de adecuación, es importante enfatizar que el objetivo debe estar en no socavar "significativamente" el funcionamiento. En este sentido, debería evitarse en lo posible revelar información sobre los indicadores de riesgo, así como información muy detallada acerca de la funcionalidad general del sistema y cualquier intento de explicación que dé cuenta de la lógica global del algoritmo. Para que sea significativa la explicación, por otro lado, debe informarse sobre el resultado de la inferencia y las ponderaciones, vale decir, la información relativa a cómo las entradas se convierten en salidas para el caso (*i.e.* lógicas correspondientes a la fase de análisis dentro de la etapa de implementación). Este requisito, además, para encontrarse presuntamente satisfecho, debe lograr traducir la complejidad del bucle algorítmico de los sistemas de ML en información comprensible, *i.e.* que resulte "*inteligible y de fácil acceso, con un lenguaje claro y sencillo*", como establece el artículo 12 del RGPD. Por tanto, debe superar las barreras de opacidad alfabética e intrínseca. Por ende, si puede ser demostrado con éxito que

[258] Cit. (n. 3), párrafo 6.82 de la sentencia.
[259] Ibid., párrafo 6.5 de la sentencia.

algún enfoque explicativo de las decisiones específicas permite cumplir estos dos criterios de adecuación, *a fortiori* puede justificarse completamente esta posición intermedia.

Con miras en este objetivo, una vía positiva para abordar este inacabado equilibro es a través de un tipo de explicación basado en la sensibilidad: las explicaciones contrafácticas. Este enfoque, propuesto por WACHTER *et al* para dar lugar al cumplimiento del RGPD, se basa en la representación de un "mundo lo más cercano posible", en que se muestra cómo se alterarían los datos de entrada lo menos posible para que la clasificación sea diferente[260]. De este modo, las explicaciones contrafácticas normalmente se expresarían de la siguiente forma: "El resultado de la clasificación es X. Si un pequeño subconjunto de características hubiera sido diferente, la predicción habría sido, en cambio, Y"[261]. Aunque con algunos reparos, los contrafactuales permitirían alcanzar el anhelado punto de equilibrio, según los ejes cardinales indicados con anterioridad.

En cuanto al resguardo de la funcionalidad, los contrafácticos son, en su forma pura, locales[262], y sus conclusiones no permiten la generalización, por lo que ciertamente facilitan este objetivo[263]. Ahora bien, aún es el caso que se está revelando al menos un *proxy* de la conducta que se intenta predecir, de modo que es posible aducir que un potencial infractor, al tomar conocimiento sobre ellos, podrá evitarlos en el futuro para alterar la clasificación adversa. Asumiendo esta posibilidad, en línea con lo que indican SOKOL y FLACH sobre la existencia de una muy difusa línea entre este tipo de explicación y los "ejemplos adversos", cuyo uso incorrecto podría ser la puerta de entrada para el juego, es crucial mantener esta línea roja al margen, revelando la menor cantidad de información posible[264].

Una vía para mitigar este problema sería la de restringir la información que se entrega al tipo de datos que podrían ser alterados mínimamente para obtener una calificación diversa, en lugar de proporcionar información concreta sobre qué datos en específico podrían revertir el resultado. De este modo, no se estarían revelando los indicadores ni el peso atribuible a una variable determinada, mientras que sí tendría un efecto indiciario sobre el peso de las "causas últimas" de la decisión. En última instancia, permitiría al receptor toma conocimiento de que su clasificación "no ha sido basada" en algún criterio de carácter sensible, en el sentido del artículo 22 apartado 4 del RGPD, en combinación con el artículo 9, apartado 2.

Notemos que este último planteamiento resulta mucho más significativo para el sujeto de datos que lo indicado por TODOLÍ, en el sentido de bastaría publicar qué tipos de datos emplea el algoritmo, con el fin de que los individuos puedan verificar que no usa información sensible o discriminatoria[265]. Lo anterior, pues circunscribiéndose al ámbito de la seguridad social, habitualmente concurriría la causal del artículo 9, apartado 2, letra b), la cual autoriza el uso de "*categorías especiales de datos*" (*i.e.* sobre salud, afiliación sindical, datos biométricos, etc.) en el tratamiento de datos para "*cumplir con las obligaciones*" específicas del "*responsable del tratamiento*"

[260] WACHTER, Sandra - MITTELSTADT, Brent - RUSSELL, Chris, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,* en *Harvard Journal of Law & Technology* 31 (2018) 2, pp. 3-7.
[261] SOKOL, Kacper - FLACH, Peter, *Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety*, en *SafeAI@ AAAI* (2019), p 1. [Visible en: http://ceur-ws.org/Vol-2301/paper_20.pdf] [Consultado por última vez: 9/08/2021].
[262] Son locales en el sentido de que la explicación se basa únicamente en la región que rodea al conjunto de datos y no en la funcionalidad general del sistema.
[263] SOKOL - FLACH, cit. (n. 260), pp. 1-3.
[264] Ibid.
[265] TODOLÍ, cit. (n. 255), p. 334.

(…) "*en el ámbito del Derecho laboral y la seguridad y asistencia social*". Por tanto, informar sobre los tipos de datos que usa el algoritmo, en el sentido aludido por TODOLÍ, podría invitar a la confusión de creer que el resultado de la clasificación se ha basado en una de las categorías especiales, cuyo uso se encontraría presuntamente autorizado. No obstante, recordemos que, pese a la autorización de su uso, el artículo 22, apartado 4, prohíbe que la decisión automatizada se base en las categorías protegidas por el artículo 9 del RGPD. Por ende, lo relevante es tener conocimiento acerca de qué tipos de datos han sido la causa de la decisión, y no así qué tipos de datos se emplean. En consecuencia, a diferencia de otros enfoques que se han propuesto, a través de los contrafácticos —aun informando únicamente sobre los tipos de datos necesarios para revertir la clasificación— se consigue un efecto indiciario significativo para comprender la lógica interna de la clasificación, generando, además, confianza hacia el sistema[266].

Por otra parte, los contrafácticos son evidentemente comprensibles para el público no experto, lo que permite superar la barrera del alfabetismo técnico[267]. Además, se trata de un método de explicación "sin abrir la caja negra", por lo que —a diferencia de otros métodos explicativos— es posible proporcionarlas aún en los casos en que se emplean algoritmos intrínsecamente opacos, como es el caso de ciertos modelos de DL[268]. En otras palabras, es posible prescindir de la interpretabilidad del modelo para proporcionar la explicación. Por ende, cumplen con el segundo criterio de adecuación mencionado.

También es cierto que algunos autores han acusado que los métodos explicativos contrafactuales adolecen de una superficialidad insalvable, por lo que volverían banal el contenido de la explicación que deriva del derecho de los sujetos de datos a conocer la lógica interna aplicada, el cual sería mucho más "profundo" que un simple contrafáctico[269]. Las críticas, en general, apuntan a que "no proporcionan una visión global" de la interacción de las demás variables para el caso específico[270]. Así, no sería posible para el individuo visualizar "el peso atribuible a cada variable", impidiendo en última instancia detectar casos injustos o discriminatorios[271].

Nada obsta, por supuesto, que esta característica considerada problemática para otros presupuestos fácticos podamos considerarla como algo conveniente para nuestro caso. Lo anterior, pues se ha tratado no de vislumbrar el método óptimo para cumplir con el mayor alcance del derecho a la explicación posible, sino de conciliar su cumplimiento —a través de un contenido informacional deliberadamente restringido— con los intereses que derivan de la funcionalidad. Después de todo, la alternativa aparentemente más próxima sería, con miras a la funcionalidad, un grado de opacidad intolerable que traería como consecuencia la prohibición de facto[272].

---

[266] En esta línea, se ha dado cuenta en la literatura sobre cómo los contrafácticos podrían ayudar incluso a detectar el impacto dispar. Vid. ZAFAR, Muhammad - VALERA, Isabel - RODRÍGUEZ, Manuel - GUMMADI, Krishna, *Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment*, en *Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee* 1 (2017), pp. 1117-1180.

[267] Vid. MILLER, Tim, *Explanation in artificial intelligence: Insights from the social sciences*, en *Artificial intelligence* 267 (2019), pp. 1-38.

[268] WACHTER *et al*, cit. (n. 259), p. 3.

[269] KAMINSKI, Margot, *The right to explanation, explained,* en *Berkeley Tech. LJ* 34 (2019) 189, p. 217.

[270] TABARRINI, cit. (n. 171), pp. 17-19.

[271] Ibid.

[272] TODOLÍ, cit. (n. 255) p. 329.

Así, este enfoque basado en una política de transparencia en dos niveles distintos permite articular de manera más coherente los valores en competencia, ajustándose a las expectativas democráticas y a los instrumentos que el RGPD proporciona. No obstante, debemos ser cuidadosos con extrapolar a otros supuestos las conclusiones obtenidas a partir del caso SyRI. Para la búsqueda de un balance óptimo, cada caso requerirá una justificación detallada y precisa a la luz de los diversos factores contextuales que se presenten. En este sentido, el planteamiento que se ha presentado es válido únicamente desde la óptica del derecho fundamental a la protección de datos, pues tal como ha planteado el tribunal en su sentencia, existe por parte del Estado una "responsabilidad especial al aplicar nuevas tecnologías" para equilibrar los beneficios con el grado de injerencia en los derechos. Más aun, como indica TODOLÍ, no se trata de la aplicación de perfiles a empresas de elite destinadas a cometer grandes fraudes, sino de "colectivos vulnerables"; generalmente "con incapacidades permanentes, invalidez o en desempleo"[273].

Pero la posición aquí defendida también encuentra una justificación adicional. La era del *Big Data* no ha hecho más que comenzar, y —como anticiparon hace más de un decenio DE HERT y GUTWIRTH— los seres humanos nos hemos vuelto cada vez más "detectables, rastreables y correlacionables, mucho más allá de nuestro control"[274]. Por lo tanto, la necesidad más urgente y prioritaria en el contexto sociotécnico actual y futuro, antes incluso que oponerse al perfilamiento, es articular medidas que permitan someter a márgenes de racionalidad su ejercicio. Es así como, dicho sea de paso, el elemento vertebrador de la argumentación que se ha presentado gravita en no cómo frenar la elaboración automatizada de perfiles, sino en proporcionar una interpretación alternativa a la prohibición que intente compatibilizar su uso con los derechos fundamentales de los sujetos de datos.

Seguro que siguiendo esta lógica se crearán algunas ineficiencias en el camino —ineludiblemente—, pero ante el evidente desequilibrio de poder que genera el uso de *Big Data*, puede que cierto grado de ineficiencia quizás deba ser visto como una garantía, y no necesariamente como un obstáculo.

CONCLUSIONES

El presente estudio confirma la hipótesis planteada inicialmente, demostrando que la opacidad estratégica en la elaboración automatizada de perfiles de riesgo puede tener efectos relevantes sobre el derecho fundamental a la protección de datos personales, particularmente cuando no existen suficientes garantías compensatorias para mitigar tales efectos. A través de un análisis detallado de las técnicas de ML y el examen de los elementos esenciales de la transparencia algorítmica, se ha demostrado la complejidad y la necesidad de equilibrio entre los valores en competencia.

Este análisis se concretó mediante el caso de estudio "SyRI", que ha permitido evaluar cómo se aplican y limitan en la práctica el principio de transparencia y las garantías del RGPD. El análisis jurisprudencial de este caso evidencia la tensión entre la transparencia y la eficacia operativa, y pone de manifiesto la necesidad de abordar estos temas con un enfoque integral que incluya aspectos técnicos, legales y éticos.

[273] Ibid.
[274] GUTWIRTH - DE HERT, cit. (n. 103), p. 291.

El estudio subraya que los algoritmos de ML, por su naturaleza, pueden reflejar y perpetuar los sesgos y patrones de discriminación inherentes a nuestra sociedad. Esta preocupación se intensifica en los casos en que se procesan datos de grupos vulnerables. Por tanto, resulta esencial que los algoritmos empleados en estos contextos de alto riesgo sean transparentes y examinables, lo que permitiría detectar y corregir conexiones espurias, sesgos o patrones de discriminación en un sentido jurídicamente problemático.

Considerando lo anterior, uno de los derechos más relevantes que consagra el RGPD es el de acceso a la lógica interna del tratamiento de los datos, cuando se toman decisiones automatizadas basadas en datos personales, incluida la elaboración automatizada de perfiles. Este derecho no requiere un entendimiento detallado de todas las operaciones técnicas del algoritmo por parte del titular, sino una explicación simplificada de cómo se ha tomado una decisión basada en los datos del individuo. Dicha explicación debe ser fácil de entender, accesible, y explicitada en un lenguaje claro y simple.

En esta línea, superar las barreras de opacidad algorítmica es un desafío clave. La opacidad puede ser intencional, como en los casos donde la divulgación podría comprometer la funcionalidad del sistema, la ventaja competitiva o incluso derechos de terceros. Asimismo, la opacidad puede ser también alfabética o intrínseca, como cuando las explicaciones son demasiado técnicas para ser comprendidas por un público no especializado, o cuando los modelos subyacentes son tan complejos que incluso los desarrolladores pueden tener dificultades para explicar su funcionamiento.

Los contrafactuales se presentan como un posible enfoque para resolver estas barreras. Esta propuesta permite que se proporcionen explicaciones comprensibles del comportamiento del algoritmo, basadas en los registros de entrada y la sensibilidad del modelo, sin comprometer su eficacia operativa. Sin embargo, es vital no limitarse a una sola técnica y considerar una variedad de posibles soluciones, incluyendo otros métodos, como la descomposición, que podrían aportar perspectivas valiosas y complementarias.

Por otro lado, además de asegurar la comprensión individual de las decisiones basadas en algoritmos de ML, es crucial establecer medidas de transparencia que posibiliten un control experto del tratamiento de datos. Estas medidas deberían comprender Evaluaciones de Impacto sobre la Protección de Datos (EIPD) y auditorías algorítmicas regulares. Las auditorías deben ser, si es posible, llevadas a cabo por entidades externas independientes para asegurar un examen imparcial y riguroso.

Así, pese al gran avance que representa el RGPD en términos de regulación sobre la protección de datos personales y su explícita consideración de la elaboración automatizada de perfiles, este trabajo demuestra que la opacidad algorítmica sigue siendo un desafío significativo que afecta a los derechos y libertades fundamentales de las personas. Esta situación pone de manifiesto la necesidad de implementar otras medidas técnicas y organizativas adicionales para garantizar un equilibrio entre la eficacia de la detección de fraudes y la protección de datos personales.

En este orden de ideas, se destaca la necesidad de abordar la comprensión de la transparencia como un fenómeno relacional, atendiendo a los factores contextuales de las distintas categorías de receptores posibles, y desde cada una de las etapas del procedimiento automatizado, que va desde la recopilación de datos hasta la toma de decisiones posterior. Este enfoque integral resalta

la dificultad de garantizar la transparencia algorítmica, poniendo de relieve la necesidad de una evaluación continua y adaptable para equilibrar los intereses involucrados en cada caso.

Por último, esta contribución concluye que, aunque la transparencia es un componente necesario para la gobernanza algorítmica, no es suficiente por sí sola. Es imprescindible adoptar un enfoque más integral que combine transparencia, equidad, capacidad de control y consideración de los contextos sociales en los que se aplican estas tecnologías, para garantizar tanto la eficacia en la detección del fraude como la protección de las libertades y derechos de las personas. Además, a partir del análisis del caso SyRI, se puede concluir que el RGPD establece una serie de mecanismos de protección que podrían mitigar los efectos de la opacidad algorítmica, pero es fundamental continuar con el análisis y la adaptación de estos mecanismos en respuesta a los rápidos cambios en el paradigma tecnológico.

BIBLIOGRAFÍA

ANANNY, Mike - CRAWFORd, Kate, *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,* en *new media & society* 20 (2018) 3.

BAMBAUER, Jane - ZARSKY, Tal, *The algorithm game,* en *Notre Dame L. Rev.* 1 (2018) 94.

BAROCAS, Solon - SELBST, Andrew, *Big data's disparate impact,* en *California Law Review* 104 (2016) 3.

BOIX, Andrés, *Los algoritmos son reglamentos: la necesidad de extender las garantías propias de las normas reglamentarias a los programas empleados por la administración para la adopción de decisiones,* en *Revista de Derecho Público: Teoría y Método* 1 (2020). [Visible en internet: https://doi.org/10.37417/RPD/vol_1_2020_33].

BOSCO, Francesca - CREEMERS, Niklas - FERRARIS, Valeria - GUAGNIN, Daniel - KOOPS, Bert-Jaap*, Profiling technologies and fundamental rights and values: regulatory challenges and perspectives from European Data Protection Authorities,* en GUTWIRTH, Serge - LEENES, Ronald - DE HERT, Paul (editores), *Reforming European data protection law* (Dordrecht, Springer-Verlag Berlin Heidelberg, 2015).

BURRELL, Jena, *How the machine 'thinks': Understanding opacity in machine learning algorithms*, en *Big Data & Society* 3 (2016) 1.

BROUSSARD, Meredith, *Artificial Unintelligence: How Computers Misunderstand the World* (MIT Press, 2019).

BÜCHI, Moritz - FOSCH-VILLARONGA, Edward - LUTZ, Christoph - TAMO-LARRIEUX, Aurelia - VELIDI, Shruthi - VILJORN, Salomé, *Chilling effects of profiling activities: Mapping the issues,* en *Computer Law & Security Review* 36 (2020).

CHOULDECHOVA, Alexandra, *Fair prediction with disparate impact: A study of bias in recidivism prediction instruments*, en *Big Data* 5 (2017) 2.

DE LAAT, Paul, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?*, en *Philosophy & Technology* 31 (2018) 4.

EDWARDS, Lilian - VEALE, Michael, *Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for*, en *Duke L. & Tech* 16 (2017) 18.

FELZMANN, Heike - FOSCH-VILLARONGA, Edward - LUTZ, Christoph - LARRIEUX-TAMO, Aurelia, *Robots and transparency: The multiple dimensions of transparency in the context of robot technologies* en *IEEE Robotics & Automation Magazine* 26 (2019) 2.

Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns, en *Big Data & Society* 6 (2019) 1.

GARRIGA, Ana, *La elaboración de perfiles y su impacto en los derechos fundamentales: una primera aproximación a su regulación en el reglamento general de protección de datos de la Unión Europea*, en *Derechos y Libertades* 38 (2018) 2.

GOODMAN, Bryce, *A step towards accountable algorithms?: Algorithmic discrimination and the European Union general data protection*, en *29th Conference on Neural Information Processing Systems,* (2016) Barcelona. NIPS Foundation.

GOODMAN, Bryce - FLAXMAN, Seth, *European Union regulations on algorithmic decision-making and a "right to explanation*", en *AI magazine* 38 (2017) 3.

GREER, Steven, *The exceptions to Articles 8 to 11 of the European Convention on Human Rights*, en *Human rights files No 15. Council of Europe Publishing* (1997).

Grupo de trabajo sobre protección de datos del artículo 29, "Directrices sobre decisiones individuales automatizadas y elaboración de perfiles a los efectos del Reglamento 2016/679" (2018). [Visible en: https://www.aepd.es/sites/default/files/2019-12/wp251rev01-es.pdf] [Consultado por última vez: 9/08/2021].

GUTWIRTH, Serge - DE HERT, Paul, *Regulating profiling in a democratic constitutional state*, en HILDEBRANDT, Mireille - GUTWIRTH, Serge (editores), *Profiling the European citizen* (Springer, Dordrecht, 2008).

HILDEBRANDT, Mireille, *Defining profiling: a new type of knowledge*, en HILDEBRANDT, Mireille - GUTWIRTH, Serge (editores), *Profiling the European citizen* (Springer, Dordrecht, 2008).

Profiling and AML, en RANNENBERG, Kai - ROYER, Denis - DEUKER, André (editores), *The Future of Identity the Information Society. Challenges and Opportunities* (Berlín, Springer-Verlag Berlin Heidelberg, 2009).

The dawn of a critical transparency right for the profiling era, en BUS, Jacques - HILDEBRANDT, Mireille (ditors), *Digital enlightenment yearbook* (Amsterdam, 2012).

HUESO, Lorenzo, *Hacia la transparencia 4.0, el uso de la inteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales,* en *Repensando la Administración digital y la innovación pública* (Instituto Nacional de Administración Pública, INAP, 2021). [Visible en:

https://www.researchgate.net/profile/Lorenzo-Hueso/publication/349591035_Hacia_la_transparencia_40_el_uso_de_la_inteligencia_artificial_y_big_data_para_la_lucha_contra_el_fraude_y_la_corrupcion_y_las_muchas_exigencias_constitucionales/links/603799f3299bf1cc26edcaef/Hacia-la-transparencia-40-el-uso-de-la-inteligencia-artificial-y-big-data-para-la-lucha-contra-el-fraude-y-la-corrupcion-y-las-muchas-exigencias-constitucionales.pdf] [Consultado por última vez: 9/08/2021].

"*SyRI, ¿a quién sanciono?" Garantías frente al uso de inteligencia artificial y decisiones automatizadas en el sector público y la sentencia holandesa de febrero de 2020,* en *La Ley Privacidad* 4 (2020).

KAMINSKI, Margot, *The right to explanation, explained,* en *Berkeley Tech LJ* 34 (2019) 1.

LAZCOZ, Guillermo, *Modelos algorítmicos, sesgos y discriminación, ponencia presentada en IX Fórum de expertos y jóvenes investigadores,* en *Derecho y Nuevas Tecnologías* (2020). [Visible en: https://www.researchgate.net/publication/338622994_Modelos_algoritmicos_sesgos_y_discriminacion] [Consultado por última vez: 9/08/2021].

LAZCOZ, Guillermo - PARRILLA, José, *Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI*, en *Revista chilena de derecho y tecnología* 9 (2020) 1.

LEPRI, Bruno - OLIVER, Nuria - LETOUZÉ, Emmanuel - PENTLAND, Alex - VINCK, Patrick, *Fair, transparent, and accountable algorithmic decision-making processes,* en *Philosophy & Technology* 31 (2018) 4.

MARCUS, Gary, *Deep learning: A critical appraisal* (2018). [Visible en: https://arxiv.org/abs/1801.00631] [Consultado por última vez: 9/08/2021].

MILLER, Tim, *Explanation in artificial intelligence: Insights from the social sciences*, en *Artificial intelligence* 267 (2019).

O`NEIL, Cathy, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Nueva York, 2016).

Organización de Naciones Unidas, *"Informe del relator especial sobre la pobreza extrema y los derechos humanos"* 74 período de sesiones, punto 72 (b) del orden del día provisional, A/74/48.037 del 11 de octubre de 2019". [Visible en: https://undocs.org/pdf?symbol=es/A/74/493] [Consultado por última vez: 9/08/2021].

OSWALD, Marion, *Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality*, en *Information & Communications Technology Law* 27 (2018) 2.

PARRILLA, José, *La Elaboración de Perfiles Políticos en España tras la Sentencia del Tribunal Constitucional de 22 de Mayo de 2019 por la que se Declara Inconstitucional el Art. 58 bis.1 de la Ley Orgánica del Régimen Electoral General - Perfilado político en España tras el pronunciamiento del Tribunal Constitucional de España de 22 de mayo de 2019 (¿Legalizar las prácticas de Cambrigde Analytica?)* (2020). [Visible en: https://www.researchgate.net/publication/339376683_La_Elaboracion_de_Perfiles_Politicos_en_Espana_tras_la_Sentencia_del_Tribunal_Constitucional_de_22_de_Mayo_de_2019_por_la_que_se_Declara_Inconstitucional_el_Art_58_bis1_de_la_Ley_Organica_del_Regimen_] [Consultado por última vez: 9/08/2021].

POWLES, Julia - SELBST, Andrew, *Meaningful Information and the Right to Explanation*, en *Conference on Fairness, Accountability and Transparency* 7 (2018) 4.

RUDIN, Cynthia, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,* en *Nature Machine Intelligence* 1 (2019) 5.

SOKOL, Kacper - FLACH, Peter, *Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety*, en *SafeAI@ AAAI* (2019). [Visible en: https://www.semanticscholar.org/paper/Counterfactual-Explanations-of-Machine-Learning-and-Sokol-Flach/b4dd607393e8c29b4e1a6d0f1d063aa2c59889c2] [Consultado por última vez: 9/08/2021].

SCHAUER, Frederick, *Fear, risk and the First Amendment: Unraveling the chilling effect*, en *BUL rev.* 58 (1978) 685.

SCHERMER, Bart, *The limits of privacy in automated profiling and data mining*, en *Computer Law & Security Review* 27 (2011) 1.

SCHETS, Femmie, *Should I be unconsciously afraid of myself? The (dis)proportionate use of risk profiling practices by public authorities to combat social security fraud* (Tesis de Magister, Instituto de Derecho, Tecnología y Sociedad de Tilburg TILT, 2019). [Visible en: http://arno.uvt.nl/show.cgi?fid=150066] [Consultado por última vez: 9/08/2021].

STOHL, Cynthia - STOHL, Michael - LEONARDI, Paul, *Managing opacity: Information visibility and the paradox of transparency in the digital age*, en *The digital age. International Journal of Communication Systems International Journal of Communication* 10 (2016) 15.

TABARRINI, Camilla, *Understanding the Big Mind. Does the GDPR Bridge the Human-Machine Intelligibility Gap?,* en *Forthcoming, EuCML - Journal of European Consumer and Market Law* 9 (2019) 4.

TODOLÍ, Adrián, *Retos legales del uso del big data en la selección de sujetos a investigar por la Inspección de Trabajo y de la Seguridad Social,* en *Revista Galega de Administración Pública* 59 (2020) 1.

VAN DIJCK, José. *Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology*, en *Surveillance & society* 12 (2014) 2.

VAN HOECKE, Mark, *Methodologies of legal Research*, en VAN HOECKE, Mark (editor), *European Academy of Legal Theory Series* 9 (2011).

VAN DALEN, Steven - GILDER, Alexander - HOOYDONK, Eric – PONSEN, Mark, *System risk indication: An assessment of the Dutch anti-fraud system in the context of data protection and profiling* (2016) Universidad de Utrech. [Visible en: https://n9.cl/yfce6] [Consultado por última vez: 9/08/2021].

VAN SCHENDEL, Sascha, *The challenges of risk profiling used by law enforcement: Examining the cases of COMPAS and SyRI*, en REINS, Leonie (editor), *Regulating New Technologies in Uncertain Times* (The Hague, 2019).

WACHTER, Sandra - MITTELSTADT, Brent - RUSSELL, Chris, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,* en *Harvard Journal of Law & Technology* 31 (2018) 2.

WACHTER, Sandra - MITTELSTADT, Brent - FLORDI, Luciano, *Why a right to explanation of automated decision-making does not exist in the general data protection regulation,* en *International Data Privacy Law* 7 (2017) 2.

ZAFAR, Muhammad - VALERA, Isabel - RODRIGUEZ, Manuel - GUMMADI, Krishna, *Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment*, en *Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee* 1 (2017).

ZARSKY, Tal, *Transparent Predictions,* en *Revista de derecho de la Universidad de Illinois* 4 (2013) (1503-1570). [Visible en: https://ssrn.com/abstract=2324240] [Consultado por última vez: 9/08/2021].

ŽLIOBAITĖ, Indrė - CUSTERS, Bart, *Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models,* en *Artificial Intelligence and Law* 24 (2016) 2.