Revista de Derecho de la Seguridad Social

Laborum

Economía y Sociología de la Seguridad Social y del Estado Social
nº 45 (4º Trimestre 2025)
ISSN: 2386-7191 – ISSNe: 2387-0370
Págs. 323-347

¿Sistemas de Inteligencia Artificial en las Entidades Gestoras de la Seguridad Social como perpetuación de sesgos y multiplicación de efectos discriminatorios?

Artificial Intelligence Systems in Social Security Management Entities as a perpetuation of biases and multiplication of effects?

Ainhoa Nieto Garrote *Investigadora predoctoral*
Universidad de Deusto
https://orcid.org/0009-0009-7156-2936

Sumario

Cita Sugerida: NIETO GARROTE, A.: «¿Sistemas de Inteligencia Artificial en las Entidades Gestoras de la Seguridad Social como perpetuación de sesgos y multiplicación de efectos discriminatorios?». *Revista de Derecho de la Seguridad Social, Laborum*. 45 (2025): 323-347.

Resumen

El empleo de sistemas de Inteligencia Artificial por parte de las Entidades Gestoras de la Seguridad Social como vía para la mejora de la gestión de las prestaciones y la lucha contra el fraude es una realidad que no ha sido suficientemente tratada desde el punto de vista de la perpetuación de los patrones discriminatorios en la analítica avanzada y los modelos predictivos. La inquietante falta de transparencia y la posible inobservancia del Reglamento de Inteligencia Artificial y la Ley 15/2022 motivan el objeto del presente estudio.

Palabras clave

Entidades Gestoras de la Seguridad Social; analítica avanzada; sesgos; patrones discriminatorios; Reglamento de IA; sistemas de alto riesgo

Abstract

The use of AI systems by Social Security Managing Entities as a way of improving management benefits and combating fraud is a reality that has not been sufficiently addressed in terms of the perpetuation of discriminatory patterns in advanced analytics and predictive models. The worrying lack of transparency and potential non-compliance with the AI Regulation and Law 15/2022 are the motivations for this study.

Keywords

Social Security Managing Entities; advanced analytics; biases; discriminatory patterns; AI Regulation; high risk systems

1. LAS PRÁCTICAS ABUSIVAS Y FRAUDULENTAS EN DETRIMENTO DE LA SOSTENIBILIDAD DEL SISTEMA COMO DETONANTE

A nadie se le escapa que vivimos en una sociedad en la que las maniobras que persiguen el beneficio propio en perjuicio del bien común se asocian a la astucia y lejos de ser rechazadas propician escenarios poco serios que distan de los principios vertebradores de nuestro sistema de Seguridad Social. El egoísmo e individualismo que imperan en nuestra sociedad no son fenómenos banales y dañan profundamente la acción protectora de la Seguridad Social.

No obstante, lo que no estamos observando suficientemente son los efectos discriminatorios que derivan de un uso descontrolado y opaco de las herramientas de Inteligencia Artificial (en adelante, IA). Concretamente, el uso por parte del sector público y especialmente por parte de las Entidades Gestoras de la Seguridad Social.

Es una obviedad, que el fraude en la obtención de prestaciones de Seguridad Social perjudica a quienes verdaderamente necesitan protección porque supone un detrimento de la sostenibilidad del sistema, pero el mayor problema lo constituyen las diversas prácticas fraudulentas que repercuten en la recaudación, como las empresas ficticias, las que no declaran el trabajo valiéndose de la falsa autonomía o las que llevan a cabo sucesiones no declaradas, redes para el fraude organizado y otras maquinaciones ilegales mediante las que buscan evadir sus responsabilidades.

Por facilitar algunas cifras, el último informe de la ITSS que tenemos a nuestra disposición, el relativo al 2023, fija como resumen general de la investigación y detección del fraude en las prestaciones; 24.270 actuaciones inspectoras y 7.760 infracciones (3.628 de empresas y 4.132 de personas trabajadoras). En cuanto al desarrollo de los resultados de la ITSS en colaboración con el INSS, el ISM y el SEPE; 24.270 actuaciones, 1.716 empleos aflorados, 7.760 infracciones por un importe de 37.138.900,76 euros y 50.540.484,45 euros en expedientes liquidatorios de cuotas.

Dicho informe también nos ofrece resultados sobre lo aflorado fruto de otras campañas de diversa índole, como las que tienen por objeto la detección de empresas ficticias que dan de alta a personas trabajadoras inexistentes con objeto de acceder a prestaciones. Añadido a las colaboraciones con la TGSS para afrontar la diversidad de tipologías de fraudes que no cuentan con campañas específicas, refiere los siguientes resultados: 25.530 actuaciones, 3.766 infracciones por importe de 22.055.823,63 euros y 49.574.112,98 euros en expedientes liquidatorios de cuotas.

En cuanto a los objetivos conjuntos de ITSS e INSS, por ejemplo, respecto a otras prácticas como altas ficticias e incremento injustificado de las bases de cotización, incumplimientos documentales por parte del empresario, etc. refiere; 525 actas de infracción levantadas y 5.341.825,46 euros de expedientes liquidatorios.

En lo que al fraude mediante diversas técnicas para la obtención de la prestación por desempleo se refiere, recoge un total de 2.090 actas de infracción levantadas y 9.517.470,76 euros en expedientes liquidatorios.

Evidentemente, las detecciones correspondientes al empleo no declarado son subestimaciones. Entre otras cuestiones, porque precisamente no se tiene en cuenta el auge de los empleos atípicos como el trabajo en las plataformas en línea basadas en la web que requerirá una modernización de la ITSS que posibilite su detección en cumplimiento de la Directiva (UE) 2024/2831 del Parlamento Europeo y del Consejo, de 23 de octubre de 2024, relativa a la mejora de las condiciones laborales en el trabajo en plataformas.

Estas pinceladas cuantitativas no son más que una pequeña muestra del problema que suponen estas acciones para la sostenibilidad del sistema. No obstante, huelga decir que la magnitud del problema no justifica el afrontarlo de cualquier manera. Veámoslo.

Los sistemas automatizados basados en prejuicios contenidos en los datos de entrenamiento o adquiridos durante el desarrollo de los modelos algorítmicos han provocado la necesidad de acuñar conceptos como "sesgo algorítmico" y "discriminación algorítmica", que evidencian, entre otras cosas, que los resultados que obtienen no deberían generalizarse.

El Parlamento Europeo ya insistía en 2017 que cuando hablamos de análisis masivos de información debe distinguirse la calidad y la cantidad de los datos y que los sistemas analíticos de toma de decisiones nutridos de datos y procesos de baja calidad propician algoritmos sesgados, correlaciones falsas y errores. También advirtió sobre la subestimación de las repercusiones éticas, sociales y legales de los efectos discriminatorios y de la marginación del papel de los seres humanos en esos procesos deficientes que repercuten gravemente en la sociedad y en los colectivos en situación de vulnerabilidad[1].

Por medio del Informe del Relator Especial de la ONU sobre la extrema pobreza y los derechos humanos publicado en 2019 se advirtieron con claridad los riesgos vinculados al empleo de programas de IA en la gestión de las prestaciones, como, por ejemplo, para la evaluación de los requisitos de acceso, cálculo de las cuantías, pagos, clasificación del riesgo, detección del fraude, etc.[2]

El abanico de datos recogido por el Relator Especial fruto de las distintas investigaciones de diversos países indica que la dimensión del problema del fraude en las prestaciones desde el punto de vista individual suele exagerarse y que en ocasiones se le presta una atención desproporcionada (en comparación con la repercusión de las actuaciones fraudulentas por parte de los empleadores). En el contexto de lo que denomina "estado de bienestar digital", los límites de la intromisión y vigilancia aparecen desdibujados y generan serios problemas.

Este informe expone muchas ideas clave, como el hecho de que determinar los derechos de las personas basándonos en predicciones que se nutren de lo calificado como comportamiento de un grupo de la población provoca errores importantes. De la mano de este concepto tenemos que tener presente lo que denomina "personalización fundamental del riesgo", que se refiere a que el atribuir factores de riesgo a determinadas características y comportamientos personales provoca que se dé preferencia a las respuestas individuales en contextos en los que esa visión resulta inútil, como la desigualdad, la pobreza o la discriminación, sin tener en cuenta su naturaleza estructural y de afectación colectiva.

A su vez, expresa que la habitual opacidad sobre el funcionamiento de los sistemas provoca que resulte verdaderamente complicado exigir que los gobiernos rindan cuentas ante la ciudadanía para el control de la vulneración de derechos. Las personas, especialmente, las que quedan afectadas por un sistema de IA que determina su asistencia social deberían contar con medios adecuados para comprender y valorar las políticas que fundamentan estos sistemas.

En este sentido, también se podría añadir, que es habitual que los sistemas de IA a aplicar en el sector público sean previamente desarrollados por el sector privado, que puede ser más propenso a descuidar los objetivos de interés general que se presumen en el ámbito público y que han de garantizar

[1] Resolución del Parlamento Europeo, de 14 de marzo de 2017, sobre las implicaciones de los macrodatos en los derechos fundamentales: privacidad, protección de datos, no discriminación, seguridad y aplicación de la ley (2016/2225(INI))

[2] Alston, P. *Informe del Relator Especial sobre la extrema pobreza y los derechos humanos*. Naciones Unidas, 2019.

la protección de los colectivos en situación de vulnerabilidad. Eso no exime a la Administración de "su posición de garante".

El Relator Especial habla con una claridad necesaria en esta materia: los sesgos en el desarrollo de estas tecnologías son inevitables y juegan un papel determinante porque el sector de la IA no se desarrolla teniendo en cuenta cuestiones de diversidad como el género y la raza. Y va más allá, afirmando que quienes diseñan los sistemas de IA en general y en particular los que se aplican a los sistemas de bienestar social son predominantemente hombres blancos con recursos procedentes del Norte global y que, desgraciadamente, en muchos casos las hipótesis y decisiones que planteen sobre estos asuntos y lo que denomina "estado de bienestar digital" quedarán determinadas por su perspectiva y experiencias. Plantea que para compensar los sesgos y velar por los derechos humanos la conformación de los sistemas deberían superar auditorías y análisis exhaustivos.

Consecuentemente y como concepto de especial relevancia, cabe destacar que la clasificación del riesgo (piénsese, para la detección del fraude) y de las necesidades (piénsese, de acceso a las prestaciones) puede fortalecer desigualdades y discriminaciones.

Otra de las conclusiones principales que puede extraerse del referido informe es la alta probabilidad de que los análisis predictivos y otros modelos de IA reproduzcan y exacerben los sesgos que esconden los datos y las políticas existentes sobre la materia.

En definitiva, los distintos tipos de discriminación estructurales menoscaban gravemente el derecho a la protección social de determinados colectivos y resulta necesaria una cooperación en red para detectar los sesgos y corregirlos en pos de ese estado de bienestar digital que se pregona.

La terminología que ha de manejarse para abordar esta temática puede crear confusión, pero algunos estudios como el que se cita a continuación de la Universidad de Cambridge arrojan luz con definiciones sencillas[3]. Este define la analítica avanzada como una metodología para analizar datos basada en modelos predictivos, algoritmos de aprendizaje automático, procesamiento del lenguaje natural, aprendizaje profundo y automatización de procesos de negocio entre otros métodos estadísticos, que facilitan el tratamiento de información proveniente de muchas fuentes de datos distintas.

Es importante apreciar la diferencia entre el sistema descrito respecto a las técnicas tradicionales de análisis, ya que la analítica avanzada es capaz de generar sistemas automatizados de IA para generar información y predicciones conductuales mucho más profundas a partir de conjuntos de datos complejos.

Por medio del citado estudio se explica la versatilidad con la que pueden utilizarse estas herramientas, en especial, en la lucha contra el fraude en sus distintas vertientes. Por ejemplo, se refiere a un informe de la OCDE de 2021 que reflejaba cómo aumenta el empleo de estas herramientas por parte de las autoridades fiscales que se valen de lo que se denomina "auditoria inteligente", que supera con creces las técnicas tradicionales de minería de datos para detectar patrones anómalos en el ámbito de la evasión fiscal.

Las ventajas que supone el empleo de la IA no deben provocar que descuidemos el ocuparnos de las famosas cajas negras y de la deriva que toma el poder de los sesgos en la toma de decisiones

[3] Tan, E., et al. "Artificial intelligence and algorithmic decisions in fraud detection: An interpretive structural model". *Data & policy*, *Cambridge University Press*, 2023, vol. 5, e25.

automatizadas. Se advierte sobre el concepto de "reincidencia de caja negra" y la probabilidad de que responda a criterios discriminatorios.

Por consiguiente, el papel de las políticas es determinante en esta materia: las decisiones políticas sobre el área de aplicación de la herramienta analítica, por ejemplo, en determinadas zonas o barrios (que sirven de sustrato para que el algoritmo opere conforme prácticas discriminatorias tan burdas como la descrita, basadas en estereotipos clasistas y aporofóbicos) provocan falsos positivos (veremos ejemplos como "el caso *Syri*" y análogos).

Algo inadvertido hasta el momento es que cuando un algoritmo es "una auténtica caja negra" puede provocar que incluso los desarrolladores no puedan determinar los razonamientos predictivos que utiliza. Es decir, es un sistema que en cierta medida puede escapar del control de los propios programadores.

Consecuentemente, cabe tener en cuenta que cuanto más complejo y autónomo sea el sistema, como es el caso de los sistemas de aprendizaje profundo, más difícil resultará acreditar las dinámicas sesgadas y los consecuentes efectos discriminatorios.

Otro razonamiento valiosísimo que proporciona el citado estudio de Cambridge es que la incapacidad de facilitar a la ciudadanía una explicación de la analítica avanzada debilita la confianza en las Administraciones Públicas y puede vulnerar (podría decirse que, en todo caso lo hace) el principio de transparencia por el que han de regirse. En este sentido, expone que el empleo de la analítica avanzada para la detección del fraude por parte de la Administración Pública ha de velar por la protección de datos personales y principios del Derecho Administrativo como la limitación de la finalidad y la minimización de datos junto a los derechos de los administrados, como el derecho de acceso a una información que le permita comprender la decisión administrativa. Y lo que es más importante a efectos de lo que aquí me interesa, los algoritmos no deben responder a criterios sesgados ni ejecutar prácticas discriminatorias.

De la revisión de los estudios de otros países puede apreciarse el empleo del *Machine Learning* para la detección del fraude en el ámbito de la asistencia sanitaria[4].

Por su parte, Amnistía Internacional se está ocupando de investigar los sistemas de IA que se están implantando en el ámbito de la gestión de las prestaciones de Seguridad Social y de la lucha contra el fraude. Se observa, que la digitalización incontrolada de los sistemas de protección social conlleva riesgos para los derechos humanos y agrava las desigualdades. Mientras publicitan que constituye una vía para avanzar en la redistribución de los recursos, alcanzar la eficiencia de los sistemas administrativos y mejorar la detección del fraude, es habitual que no se dé una rendición de cuentas y que se utilicen sistemas opacos (de caja negra) que invisibilizan los razonamientos del modelo.

En definitiva, los efectos del sesgo de estos algoritmos dificultan el acceso a las prestaciones por parte de determinados colectivos en situación de vulnerabilidad y a su vez, los estigmatiza y persigue conforme a patrones arbitrarios y discriminatorios.

En este punto se considera oportuno hacer una aproximación a los casos más llamativos, investigados, entre otras, por Amnistía Internacional.

[4] Ünal, C., Erbuğa, G.S.: "Detection and Prevention of Medical Fraud using Machine Learning". *Acta Infologica*, 2024, vol. 8, nº 2, p. 100-117.

Conviene atender al informe denominado *Xenophobic Machines*[5] que surgió tras la detección del empleo de perfiles raciales en el diseño del modelo utilizado para la resolución de solicitudes de subvenciones para el cuidado infantil en Países Bajos, que tildaba de "solicitudes potencialmente fraudulentas" las que provenían de personas con determinadas características. Se calcula que decenas de miles de familias pertenecientes a minorías étnicas (a las que se les asignaba una puntuación de riesgo más alta) han sido falsamente acusadas de fraude por las autoridades tributarias neerlandesas, lo que les ha sumido en una situación de mayor precariedad por tener que hacer frente a deudas inasumibles e injustificadas. A fin de cuentas, el sistema atribuía directamente conductas ilegales a determinados colectivos.

El 5 de febrero de 2020, el Tribunal de distrito de La Haya dictó una sentencia por la que se estableció que el sistema algorítmico empleado por el Gobierno de Países Bajos para evaluar el riesgo de la concurrencia de supuestos de fraude a la Seguridad Social o Hacienda, no cumplía con los principios de transparencia y proporcionalidad e infringía el derecho a la vida privada consagrado en el artículo octavo del Convenio Europeo de Derechos Humanos.

Otro caso que cabe destacar es el de los sistemas de IA discriminatorios de la agencia de bienestar social de Suecia. Varias investigaciones revelaron que el sistema sueco seleccionaba injustamente a colectivos en situación de vulnerabilidad para someterlos a inspecciones en materia de fraude en las prestaciones sociales. Concretamente, se trata de un sistema de aprendizaje automático implementado desde 2013 que toma como referencia características como el género, el país de origen o incluso el país de origen de los progenitores, la capacidad económica o la formación. Sobre dicha información, el sistema asigna puntuaciones ante presuntos riesgos de llevar a cabo actuaciones fraudulentas, de forma que las personas que obtienen puntuaciones más altas conforme a los cálculos del algoritmo son las que resultan objeto de investigación. Y lo que es peor, a dichos solicitantes se les atribuye arbitrariamente una "intención dolosa" desde ese momento inicial. Las dinámicas del funcionamiento generan una retroalimentación de la información que obtiene el algoritmo, de forma que el sesgo queda arraigado y los efectos discriminatorios se multiplican.

En Francia existe otro ejemplo: el sistema algorítmico de puntuación de riesgos utilizado por la Caja Nacional de Prestaciones Familiares (CNAF). Amnistía Internacional, junto a 14 entidades asociadas a la coalición que dirige La Quadrature Du Net (LQDN) interpusieron una demanda ante el máximo tribunal administrativo de Francia por la que solicitan el cese definitivo del empleo del modelo de puntuación de riesgos del sistema algorítmico utilizado por el sistema de seguridad social francés. Concretamente, este modelo se basa en asignar puntuaciones para calificar a personas que según sus criterios puedan estar llevando a cabo actividades fraudulentas en orden a las prestaciones. De esta forma, el algoritmo asigna a cada persona beneficiaria de prestaciones "familiares y de vivienda" puntuaciones en el rango de entre cero y uno. Cuanto más se acerque el resultado a uno, mayor probabilidad de fraude se asocia a la persona. Una vez más, el razonamiento se sesga con dinámicas discriminatorias.

En Dinamarca también se han detectado prácticas de vigilancia masiva y monitoreo de las familias solicitantes y beneficiarias de prestaciones sociales. En este caso, su sistema de bienestar social estaba gestionado mediante herramientas de IA que centran el foco en las personas más desprotegidas, ya sean personas con discapacidad, las que tienen bajos ingresos, las migradas, las solicitantes de asilo, las refugiadas, etc. Se detectaron hasta sesenta modelos algorítmicos en teoría diseñados para detectar fraude en las prestaciones sociales, pero materialmente constituidos como una vía para perseguir a las personas históricamente marginadas por motivo de su situación económica,

[5] Amnistía Internacional: *Injusticia codificada. vigilancia y discriminación en el estado de bienestar automatizado de Dinamarca*. Londres, Amnesty International Ltd., 2024.

origen, etc. También se ha conocido el empleo de un sistema discriminatorio establecido en Serbia, que generó nuevas barreras en el acceso a la protección social.

En Polonia se declaró inconstitucional un sistema que determinaba el acceso al subsidio por desempleo y en qué cuantía[6]. *Contrario sensu*, en Austria sigue aplicándose un sistema algorítmico que categoriza a las personas desempleadas determinando su cobertura.

En este punto procede traer a colación el estudio de Martín López[7], que pone el foco en los patrones discriminatorios, pero desde el punto de vista de los sistemas de IA aplicados al ámbito de la Inspección Tributaria. Concretamente, analiza si la elaboración de perfiles de riesgo y patrones de incumplimiento junto al modelo predictivo basado en *Machine Learning* puede derivar en actuaciones discriminatorias.

Este autor hila fino en la conceptualización de los sesgos que pueden adquirir los sistemas de IA ya que es importante diferenciar entre los sesgos que proceden de los datos seleccionados o aquellos que forman parte del diseño del propio modelo.

El primer escenario se relaciona directamente con el sesgo en el muestreo y sus efectos se traducen en que la infra o la sobre representación de determinado colectivo provoca que el modelo replique la estructura, que, a su vez, se materializa en el resultado desviado.

La segunda parte se podría denominar "efecto multiplicador", porque el modelo no se limita a repetir los resultados (lo que ya sería negativo) sino que los amplifica y perpetúa porque se retroalimenta de ellos. Resultados, que derivan de cómo ha sido entrenado el algoritmo. Cuanto más se acota el enfoque más se refuerza el sistema que señalará sistemáticamente a las personas con las características predefinidas. El referido estudio lo ejemplifica con el análisis de las personas obligadas tributarias e investigadas y, además, sancionadas. Además de denotar un prejuicio bastante evidente sobre la reincidencia, cuanto más se investigue a ese colectivo, más se centrará el sistema en él alejándose del resto.

La exposición del autor contiene otros planteamientos que, considero, también merecen una respuesta expresa.

Resulta esencial distinguir con claridad el objetivo de las discusiones que planteamos sobre esta materia, ya que, el cuestionar la forma en la que vienen utilizándose los sistemas de IA por parte de la Administración Pública no persigue que se descarte el empleo de estos sistemas. Sería contraproducente no valorar y utilizar las herramientas que nos ayudan a ser más eficientes. El objetivo es exigir la debida transparencia y garantías que impidan que los sesgos se perpetúen y se fortalezcan multiplicando los efectos discriminatorios directos o indirectos.

En segundo término, desde mi punto de vista es alarmante que hablemos de "discriminación jurídicamente relevante" como una especie de presupuesto de gravedad. Considero que la prevención no debe depender de la reprochabilidad jurídica de las dinámicas, que, valga la ocasión, quizás no se adecue a la realidad. Que cualquier inspección o gestión de recursos opere siguiendo patrones sesgados es socialmente relevante y reprochable y nos coloca a los operadores jurídicos en posición de exigir que se depuren los sistemas y se adopten medidas de prevención.

[6] Tribunal Supremo de Polonia, causa núm. K 53/16, 6 de junio de 2018

[7] Martín López, J.: "Inteligencia artificial, sesgos y no discriminación en el ámbito de la inspección tributaria". *Crónica Tributaria*, nº 182, 2022. Pp. 51-89.

Es cierto que como dice el autor no conviene caer en reduccionismos y no es lo que aquí se pretende, pero precisamente en cuestiones que no deberían suscitar duda alguna, como es la prevención de prácticas susceptibles de provocar situaciones de discriminación directa o indirecta, no deberíamos ser rígidos. Tenemos a nuestra disposición información suficiente para apreciar los riesgos de utilizar los sistemas de analítica avanzada predictiva como para evitar caer en discusiones teóricas que retrasen el impulso de las vías de actuación para su corrección. La realidad nos lo exige y como veremos, el legislador europeo lo tiene muy claro.

2. ¿QUÉ CONOCEMOS SOBRE LOS SISTEMAS QUE ESTÁN UTILIZANDO LAS ENTIDADES GESTORAS?

2.1. El Modelo de Priorización de Citas del INSS. ¿También *SAS Fraud Framework*?

Lo cierto es que no resulta sencillo saber qué herramientas de IA están utilizando las Entidades Gestoras de la Seguridad Social, especialmente en lo que a la IA predictiva se refiere. Algunos estudios nos aproximan a las diversas manifestaciones de la digitalización de las Entidades Gestoras[8]. Sin embargo, en esta ocasión centraré el foco en aquellas que considero más delicadas.

En este contexto, conviene atender a lo que nos transmiten quienes tienen directa relación con las Entidades Gestoras o forman parte de ellas.

Llamativamente, aunque no dispongamos de suficientes fuentes oficiales relativas al uso de sistemas de IA por parte de las Entidades Gestoras de la Seguridad Social, en un breve monográfico publicado en el año 2018 del entonces responsable de Área de Estadísticas y Análisis de Datos (Gerencia Informática de la Seguridad Social) ya se hacía referencia a la aplicación de técnicas de aprendizaje automático o *Machine Learning* y el análisis predictivo como forma de optimización de los procesos de negocio y el incremento de la productividad de las unidades administrativas, además de la lucha contra el fraude en el cobro de las prestaciones[9].

Escudero Rivas[10], Director Gerente de la Gerencia de Informática de la Seguridad Social, define el análisis predictivo como una técnica que aúna datos históricos y modelos estadísticos con objeto de predecir sucesos futuros o los resultados más probables. En esencia, el objetivo que fija esta técnica es la conformación de patrones que faciliten predecir acontecimientos futuros.

Esta primera definición ya debería provocarnos una llamada de atención en relación con la salvaguarda del sistema respecto a los patrones sesgados, en tanto los datos históricos no deberían darse automáticamente por admisibles para determinar decisiones futuras.

En esta línea se pronuncia el legislador europeo al advertir que los sesgos pueden ser inherentes al conjunto de datos que subyacen a los sistemas (especialmente si se utilizan datos históricos o los que derivan de la aplicación de los sistemas en entornos reales). Los resultados quedarán determinados por dichos sesgos inherentes a la información que nutre los sistemas cuyo aprendizaje automático

[8] Fernández Ramírez, M.: "Inteligencia Artificial, algoritmos predictivos y gestión tecnológica de la Seguridad Social" en *Las transformaciones de la Seguridad Social ante los retos de la era digital: VII Congreso Internacional y XX Congreso Nacional de la Asociación Española de Salud y Seguridad Social*. Murcia, Laborum, 2023. Tomo I, pp. 155-174.

[9] Pardo García, J.: "La analítica avanzada de datos en la Seguridad Social"*, Astic*, 2018, disponible en: https://www.astic.es/wp-content/uploads/2018/06/boletic82-monografico4-juanpardo.pdf

[10] Escudero Rivas, C.: "El análisis predictivo en las relaciones laborales y de seguridad social" en VV.AA.: *VII Congreso Internacional y XX Congreso Nacional de la Asociación Española de Salud y Seguridad Social. Las transformaciones de la Seguridad Social ante los retos de la era digital,* Murcia, Laborum, 2023. Tomo I, pp. 29-41.

hace que se retroalimenten y aumenten gradualmente, perpetuando y amplificando sus efectos discriminatorios, que afectan especialmente a los colectivos en situación de vulnerabilidad[11].

Seguidamente, el citado autor ejemplifica cómo puede utilizarse esta herramienta en las vicisitudes de las relaciones laborales de forma que se obtenga información para la toma de decisiones y se logre una anticipación a las situaciones o necesidades futuras. Así, se refiere a la probabilidad de que una persona trabajadora no pueda prestar sus servicios o demande servicios de la Seguridad Social, entre otras.

Aunque no lo mencione, se sobreentiende que valora positivamente dicha aplicación a efectos internos de gestión de las Entidades Gestoras. No obstante, sin perjuicio de la protección normativa y jurisprudencial en la materia, no está de más mencionar el peligro que supondría la utilización de este tipo de aplicaciones por parte de los empleadores, ya que no sería disparatado pensar que muchos de ellos las aprovecharían para tomar decisiones extintivas discriminatorias con objeto de abaratar costes.

Otra ejemplificación que diría que sin lugar a dudas entraña riesgo es la aplicación de estos sistemas en la gestión del rendimiento. Si bien el autor la relaciona con la retención del talento, pensando en la protección de las personas trabajadoras yo lo veo como una forma más de hipervigilancia y de cosificación de las mismas. Adviértase que el citado autor menciona que este sistema facilitaría conocer la probabilidad de abandono de la organización. Lo que no se menciona es que manejar este tipo de información por parte del empleador puede provocar decisiones reaccionarias como el despido.

En cualquier caso, el planteamiento de la utilización de este tipo de herramientas no puede hacerse teniendo en cuenta únicamente los aspectos positivos que, casualmente, lo son para la parte que domina la relación.

Como tres ventajas principales del empleo de esta herramienta en la Seguridad Social fijadas por la citada doctrina, aparecen: la detección del fraude, la gestión eficiente de los recursos y la mejora en la toma de decisiones.

La investigación de Fernández Orrico[12], tal y como nos explica, en parte nutrida de entrevistas directas a personas trabajadoras que conocen de primera mano el funcionamiento de las Entidades Gestoras, precisamente, por la falta de información oficial, resulta reveladora. Parece que el INSS está utilizando algoritmos predictivos, por un lado, con el objetivo de conocer los casos de IT más cercanos al momento del alta y poder así citar a las personas de forma prioritaria, y, por otro lado, con objeto de detectar actuaciones fraudulentas para la percepción de la prestación por IT.

A su vez, la programación de dichos algoritmos responde, por un lado, al análisis de los expedientes de baja por IT pendientes de revisión y, por otro lado, al análisis de los expedientes de bajas por IT ya revisados. Se indica que según el análisis comparativo se obtiene una puntuación entre cero y uno y se interpreta de forma que cuanto más cercano al uno sea el resultado, más próxima se encuentra la persona del alta médica.

[11] Considerando 67º del Reglamento de IA.

[12] Fernández Orrico, F.J.: "La IA como instrumento de predicción en materia de altas médicas. Un análisis a la vista de la normativa sobre protección de datos y la propuesta de Reglamento (UE) de IA" en: VV.AA.: *VII Congreso Internacional y XX Congreso Nacional de la Asociación Española de Salud y Seguridad Social. Las transformaciones de la Seguridad Social ante los retos de la era digital*, Murcia, Laborum, 2023. Tomo I, pp. 133-153.

Obviamente, nutrir el sistema con dichos expedientes supone que el Modelo de Priorización de Citas maneje datos de carácter personal, pero por lo que aquí me interesa no me detendré en la protección de datos, ya que quiero centrar el foco en una discusión que todavía no observo suficientemente en la doctrina: el riesgo que supone el empleo de estas herramientas concretas para perpetuar y multiplicar dinámicas discriminatorias.

Es cierto que contamos con estudios como el de Villar Cañada[13], que manifiestan los patrones discriminatorios originados por la brecha digital y sus múltiples efectos directos o indirectos. Es más, la misma autora advertía desde 2021 los riesgos que supone la búsqueda de la viabilidad de la Seguridad Social mediante la digitalización desde una mirada relacionada con la brecha digital[14]. También contempla una reflexión muy relevante a efectos del enfoque que se le da al problema mediante el presente estudio: es necesario un análisis de datos adecuado para minimizar los riesgos inherentes al empleo de los sistemas de IA. Se refiere a que no podemos limitarnos a recoger datos, sino que tenemos que analizarlos y reflexionar sobre ellos como forma de evolucionar desde la sociedad de la información hacia la del conocimiento. Añade la autora, que los sesgos algorítmicos derivan de los existentes en la sociedad y que los datos que sirven para conformar los algoritmos han de ser analizados para que no operen conforme a razonamientos discriminatorios.

Esta autora plantea acertadamente que el análisis predictivo puede facilitarnos la obtención de indicaciones o recomendaciones, pero no determinar el nivel del riesgo o la necesidad objeto de valoración.

A su reflexión, yo añadiría que incluso debemos colocarnos en una posición anterior a la consideración de que los modelos predictivos nos ofrecen indicios *per se*, ya que los resultados pueden responder a criterios sesgados.

Independientemente de que después de la obtención de los resultados del análisis predictivo, en teoría (y digo en teoría porque como bien denuncia Fernández Orrico, no contamos con suficiente transparencia por parte del INSS), se revisen todas las historias clínicas y resto de datos a tener en cuenta para citar a la persona a la revisión, parece evidente que el resultado algorítmico ya predispone a la persona que tiene que valorarlo.

2.1.1. Posible multiplicación de los efectos del sesgo de género en la gestión de la prestación por IT

Mientras persista la toma de decisiones sesgadas basadas en estereotipos es evidente que habrá sistemas de IA que funcionen sustentados por esos patrones discriminatorios. Está en nuestras manos impulsar el cambio en las personas y exigir que el empleo de IA no sea una vía para perpetuar la discriminación.

Incorporar la IA a entornos tan delicados como las Entidades Gestoras de la Seguridad Social requiere una perspectiva de género. Las fases son lógicas: la discriminación estructural existente en la sociedad se traslada al entrenamiento de las herramientas de IA mediante una nutrición cargada de estereotipos y sesgos (muestreo sesgado). La fase más dañina es la retroalimentación de esos sistemas

[13] Villar Cañada, I.M.: "El impacto tecnológico en el ámbito sociolaboral: ¿obstáculo u oportunidad para la sostenibilidad del sistema de pensiones?". *Revista de Derecho de la Seguridad Social, Laborum*. Extraordinaria nº 7 (2024): 241-261.

[14] Véase Villar Cañada, I.M.: "La digitalización y los sistemas de protección social: oportunidades y desafíos". *Revista de Trabajo y Seguridad Social. CEF*, 459, 173-205. El mismo razonamiento queda respaldado por Romero Coronado, J.: "Impacto del Big Data y la Inteligencia Artificial en la gestión de la Seguridad Social". *Revista de Derecho de la Seguridad Social, Laborum*. Extraordinaria 6, 2024, pp. 49-70.

nutridos con dicha información, que refuerzan y multiplican los efectos sea cual sea el campo donde se utilicen. Esto es, los sesgos provocan efectos tanto antes de conformarse el sistema de IA, como en cualquiera de las fases de desarrollo o en la aplicación del modelo y los resultados.

El estudio de Pérez-Ugena Coromina[15] nos ofrece una conceptualización muy acertada de la forma en la que opera la discriminación de género en los sistemas de IA. Además, plantea una vía de actuación que resulta esencial: la realización de auditorías algorítmicas y como no podía ser de otra manera, una educación que pueda propiciar una corrección de fondo y a largo plazo. Atacar el problema de raíz. No obstante, algo que no comparto, en relación con la consideración de que la aparición de sesgos en la IA radica en la discriminación indirecta, es que la discriminación indirecta no guarde relación con la intencionalidad. Sobre esto me detendré más adelante.

En este sentido, yo añadiría que los equipos que se dediquen a entrenar estas herramientas, que tienen perfiles técnicos, deberían comenzar a contar con una formación que les sensibilice sobre esta realidad y les facilite herramientas para detectar información que no debiera utilizarse para el entrenamiento de los algoritmos. A su vez, creo necesario que los equipos sean multidisciplinares y que cuenten con profesionales de las ciencias sociales que velen por depurar las malas prácticas ante la dependencia que muestran los resultados respecto a su diseño y consecuentemente, de los programadores.

Se ha acuñado el término "caja negra" para que podamos referirnos a la opacidad del funcionamiento de estos sistemas. Pero cuando hablamos de su inclusión en el sistema de Seguridad Social directamente es inadmisible que asumamos dicha falta de transparencia.

Demos un paso más. En el ámbito médico, la doctrina científica define el sesgo de género como "*la diferencia en el tratamiento de ambos sexos con un mismo diagnóstico clínico, pudiendo tener consecuencias positivas, negativas o neutras para la salud de los mismos*". El citado estudio facilita un importantísimo razonamiento sobre cómo se ha ido construyendo el sesgo de género desde la investigación y generación del conocimiento a la asistencia sanitaria[16].

Para el caso que nos ocupa, voy a destacar algunas de las apreciaciones de los sesgos de género a nivel asistencial. La citada doctrina denomina "ceguera de género" al hecho de obviar diferencias que revisten importancia. *Contrario sensu*, otro error consiste en asumir diferencias cuando no existen como consecuencia de los "estereotipos dicotómicos asociados". Este fenómeno al que de forma general se refiere como "sesgo de género cognitivo, social y constitucional", provoca creencias sesgadas, que propician prácticas discriminatorias en la prevención y promoción de la salud y la práctica clínica.

Si bien este fenómeno de los sesgos en los algoritmos y sus efectos discriminatorios puede perjudicar a cualquiera, una vez más quienes salen peor paradas son las mujeres y esta conclusión se puede extraer de la descripción de las dimensiones que adopta el sesgo de género en la asistencia sanitaria. Del listado que nos ofrece el estudio, voy a destacar algunas de las afirmaciones que me parecen más relevantes.

Califico como piedra angular, la existencia de sesgos de género en el diagnóstico. Mediante el citado estudio se afirma que es posible que algunos síntomas atípicos de una patología no se

[15] Pérez-Ugena Coromina, M.: "Sesgo de género (en IA)". *EUNOMÍA. Revista en Cultura de la Legalidad*, 2024, nº 26, pp. 311-330.

[16] Cabanillas-Montferrer, T., et al.: "El sesgo de género en la asistencia sanitaria: definición, causas y consecuencias en los pacientes". *MUSAS. Revista de Investigación En Mujer, Salud y Sociedad*, 2022, vol. 7, nº 1, pp. 106-129.

consideren y que, como consecuencia, algunas personas (en su mayoría, mujeres) queden excluidas del sistema sanitario. A su vez, eso provoca que solo sean objeto de estudio los casos de pacientes que lleguen a lo que se denomina "esfuerzo diagnóstico". Esto es, explica que el sesgo de género afecta a los esfuerzos en primer término de diagnóstico (mediante la selección de las pruebas a realizar) y en segundo término de tratamiento.

A su vez, mediante el referido estudio se detectó una diferencia de comportamiento que cabe destacar: los hombres acuden con mayor frecuencia a los servicios sanitarios especializados o de urgencia de los hospitales y las mujeres acuden con mayor frecuencia a la atención primaria. Consecuentemente, ¿las mujeres restan importancia a los síntomas y asumen una resistencia física mayor? Difiero de la afirmación sobre que esta utilización diferencial sea responsabilidad de los pacientes además del sistema sanitario, puesto que es la asunción de los roles de género favorecidos por la educación (especialmente los relacionados con el cuidado) los que colocan a la mujer en el escenario de la resistencia en lugar de la pedida de ayuda.

También se aprecia una diferencia en el tiempo de espera para recibir asistencia sanitaria. En la misma línea, no solo se observa que las mujeres tardan más en solicitar ayuda, sino lo que es más preocupante: cuando la solicitan tardan más en ser atendidas.

La doctrina científica advierte de las graves consecuencias del sesgo de género en la salud de las mujeres, que, en definitiva, pueden conllevar que la enfermedad empeore. El círculo vicioso descrito, que comienza con la diferencia en las percepciones de los síntomas fruto del papel que las mujeres desempeñan en la sociedad determina la diferencia asistencial de principio a fin (desde el diagnóstico hasta el tratamiento).

Esta realidad sigue existiendo por la falta de investigación con perspectiva de género en este ámbito, que posibilita la perpetuación del problema sin que llame la atención. Denuncia el citado estudio, que en esencia es una forma de legitimar por omisión los defectos de la sociedad trasladados a la asistencia sanitaria y que para afrontarlo es necesario que los profesionales reciban formación especializada en clave de género.

A su vez, puede observarse cómo se está insistiendo en la necesidad de incorporar la perspectiva de género a la valoración de las enfermedades profesionales para una mejora de la prevención[17].

Lo que yo planteo es trasladar esta visión a la incorporación de los sistemas de IA por parte de las Entidades Gestoras de la Seguridad Social. Insistiré en que un cauce para comenzar a trabajarlo es el análisis de impacto de género.

Como puede observarse, la disciplina médica tiene que buscar soluciones a este problema, pero desde la perspectiva jurídica no puede ignorarse. Teniendo en cuenta los razonamientos expuestos mediante esta mirada interdisciplinar que considero necesaria, procede en este punto plantear la siguiente reflexión.

Desde mi punto de vista, los sesgos de género que afectan especialmente a las mujeres en el tratamiento de sus dolencias y consecuentemente, en la valoración de la IT o posible IP, pueden perpetuarse y multiplicarse empleando sistemas de IA predictiva de forma descontrolada.

[17] Igartua Miró, M.T.: "El enfoque de género y la enfermedad profesional: reflexiones desde el ordenamiento español". *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo*, ADAPT University Press, vol., 1, nº2, 2023.

2.1.2. Extensión del razonamiento a cualquier colectivo en situación de vulnerabilidad

Como siempre, los motivos de discriminación pueden responder al origen étnico o racial, orientación sexual, estatus económico, edad, religión y un largo etcétera.

La doctrina ya contempla la inclusión de las herramientas de IA en la asistencia sanitaria como un problema de justicia social que puede provocar efectos discriminatorios hacia los distintos colectivos de personas. Categorizar y atribuir a las personas comportamientos con motivo de ciertas características es muy peligroso y los datos que conforman los algoritmos para su aplicación en la asistencia sanitaria comienzan a cuestionarse. Por ejemplo, nutrir el sistema con datos extraídos de historias clínicas originarias de un contexto en el que no existe diversidad étnica cuando se va a utilizar para estudios genómicos. Como consecuencia del sesgo en los estudios, la aplicación de las herramientas a las poblaciones minoritarias no resulta exitosa. En definitiva, mientras no se corrijan estas formas de proceder el empleo de la IA reproducirá el problema y por lo tanto los efectos discriminatorios[18].

A modo de cierre y a propósito del interrogante que añado en el título del bloque, ¿el INSS también está utilizando el *SAS Fraud Framework*?

Puede que con ánimo de pasar desapercibida por el carácter informal del canal (página web, sin ninguna elaboración documental), la única y sucinta información oficial parece que fue publicada en 2018. Es la siguiente: "*Las líneas de trabajo desplegadas por el INSS en la aplicación de las tecnologías más avanzadas de analítica avanzada han consistido en definir modelos para optimizar el control médico del INSS ante las incapacidades temporales, a través de la herramienta SAS, a modo de programa piloto en seis direcciones provinciales, permitiendo no citar a un reconocimiento médico a los casos que están correctamente justificados dentro del periodo de los primeros 365 días. En estos casos, los modelos predictivos nos indican de manera mucho más ágil cuáles son los casos susceptibles de la actuación de los inspectores médicos*"[19].

Si bien la explicación del objetivo para el que se supone que se emplea la herramienta SAS (*Software* de Análisis Estadístico) parece encajar con el sistema de priorización de citas, la duda es si también se emplea el *SAS Fraud Framework* como solución específica perteneciente a esa misma empresa (privada, por cierto) para perseguir el fraude. Veremos que la TGSS sí la utiliza.

2.2. El empleo de la IA para el análisis de las personas beneficiarias del IMV

Para hacernos a la idea de la odisea que en muchos casos ha supuesto el acceso, en caso de lograrlo, a prestaciones como el IMV conviene atender a casos reales, como el que describe la periodista Sara Mesa en su libro titulado *Silencio administrativo. La pobreza en el laberinto burocrático*[20]. Una clara muestra de la desinformación que se propaga en cuanto a las personas en situación de calle y un reflejo de las incongruencias de las exigencias desmedidas como reflejo de la distancia que media entre las cabezas pensantes de los requisitos y el colectivo de personas que necesitan ayuda. Yo lo denominaría aporofobia social y estructural.

[18] De Miguel Beriain, I.: "Medicina personalizada, algoritmos predictivos y utilización de sistemas de decisión automatizados en asistencia sanitaria: Problemas éticos". *Dilemata*, 2019, nº 30, pp. 93-109.

[19] Disponible en: https://revista.seg-social.es/-/big-data-contra-el-fraude

[20] Mesa, S.: *Silencio administrativo. La pobreza en el laberinto burocrático*. Barcelona, Editorial Anagrama, S.A.U., 8ª ed., 2019.

Remitiéndome de nuevo a lo que recoge la exposición de Escudero Rivas, explica el análisis de grupos como herramienta para obtener conjuntos de personas con características similares y llevar a cabo acciones específicas. Afirma que "*se ha seleccionado el conjunto total de beneficiarios del IMV cuyas solicitudes han sido aprobadas, incluyendo datos sociodemográficos de sus unidades de convivencia, con el objetivo de detectar características que permitan llegar a otras familias con este perfil que no han solicitado la prestación*". Y añade que "*El resultado ha ofrecido diez grupos diferenciados, que se caracterizan por tener unos atributos muy similares entre sí y muy diferentes del resto*."

Quizás el espíritu de esta técnica fuera efectivamente llegar a los supuestos susceptibles de protección, pero considero que este tipo de afirmaciones soslayan que retroalimentar el análisis con las características relativas a las personas con solicitudes aprobadas puede dilatar aún más la falta de observancia de los supuestos que debieran protegerse pero que no cuentan con la resolución favorable. A lo que hay que añadir, que los bucles de retroalimentación hacen que los nuevos datos que se obtienen como resultado de los cálculos algorítmicos sean cada vez más negativos para el colectivo que está quedando marginado.

Este ejercicio reflexivo se puede ejemplificar con algunos de los problemas que puso de manifiesto Mouvement International ATD Quart Monde (ONG inscrita en el registro de ONGs del Consejo de Europa) mediante la interposición de una Reclamación Colectiva en aplicación del Protocolo Adicional de la Carta Social Europea, frente al reino de España, por la no conformidad de diversos artículos de la Ley 19/2021, de 20 de diciembre, por la que se establece el Ingreso Mínimo Vital y de su aplicación, con los artículos 1, 7, 13, 15, 16, 17, 27, 30, 31 y E de la Carta Social Europea. Destaco esta reclamación, pero son muchas las posiciones desde las que se puso de manifiesto el mismo problema y que se mencionan por medio del citado escrito.

Las trabas que surgieron del desacierto en el diseño de la norma y la gestión de la prestación se extendieron desde la configuración de los requisitos como fuente de una gravísima inseguridad jurídica, una normalizada dilación indebida en la gestión y un abuso de la figura del silencio administrativo con la correspondiente falta de motivación. Dinámicas que logran dejar fuera de protección precisamente a supuestos de especial gravedad, a lo que hay que añadir la inclusión de restricciones subjetivas relativas a condiciones como la edad o la unidad convivencial, entre otras.

Además, cabe destacar el fenómeno de los cobros indebidos. Tres años después, en un intento de arreglar el caos provocado se consiguió el efecto contrario. Se incoaron miles de expedientes de revisión de prestaciones reconocidas que originaron un mayor empobrecimiento de las familias por no haber estudiado los supuestos con suficiente cautela antes de llevar a cabo la medida.

Pongamos el ilustrativo ejemplo que se trae a colación en la exposición del hecho primero de la referida reclamación. Una familia monomarental con dos hijos perceptora de la prestación por hijo a cargo (uno de ellos de un progenitor distinto al primero del que se está divorciando) a la que se le reconoce de oficio la prestación en aquel intento de arreglar la situación. El proceso de revisión supuso el inicio de un expediente de cobros indebidos y el cese sin notificación previa de la percepción del IMV, terminando con un requerimiento de devolución de 1.455 euros viéndose obligada, dada su situación de necesidad, a solicitar una nueva prestación. Esta última se reconoció, pero se indica que únicamente recibe 50 euros mensuales desde enero de 2022. La reclamación no ha sido resuelta lo que le ha colocado en la necesidad de judicializar el asunto como consecuencia del silencio administrativo.

En junio de 2024, cuando se interpuso la mencionada reclamación, el 73% de las solicitudes fueron denegadas por criterios de renta, patrimonio o unidad de convivencia (con un 57% de solicitantes a la espera de resolución, lo que representa alrededor de 400.000 hogares). Para entonces, únicamente se había ejecutado el 56% del presupuesto destinado a la prestación.

No pretendo hacer un análisis exhaustivo del problema de gestión del IMV, sino reflejar que automatizar con herramientas de IA la valoración de las familias susceptibles de protección mediante análisis predictivos partiendo de ineficacia del sistema, puede ser una vía para terminar de empeorar la situación y alejar todavía más del punto de mira los supuestos a los que la ayuda no llega.

Una vez más, las dificultades para alcanzar los objetivos de protección de las personas que necesitan acceder a la prestación puede ser un reflejo de la distancia entre el legislador y el colectivo de personas que sufren el problema.

2.3. Sistemas predictivos en la TGSS para la detección del fraude

He querido definir este tercer pilar porque considero que es otro ejemplo de falta de transparencia y la seriedad del asunto lo merece.

Una vez más, tenemos que acudir a fuentes oficiales pobres en cuanto al desarrollo explicativo de lo que se está haciendo, pero que caen en afirmaciones problemáticas a efectos de lo que después veremos. En un documento de la TGSS relativo a la IV Convocatoria anual de innovaciones del observatorio de innovación pública de la OCDE[21], se expone que la TGSS no disponía de medios humanos suficientes para la lucha contra el fraude mediante el manejo de herramientas de análisis y técnicas predictivas que permitan la detección de patrones de comportamiento (sostiene que para afrontar el fraude se necesita tecnología de minería de datos que permita parametrizar la información que custodia la TGSS, definir perfiles y patrones de comportamiento). Explicando, que para ello la TGSS ha tenido que adaptar sus procesos de gestión mediante nuevos recursos tecnológicos y ha conformado un equipo especializado (nada se dice que en él se incluyan a personas expertas en discriminación para su prevención) y que en colaboración con la ITSS se ha implantado la herramienta de análisis predictivo denominada *SAS Fraud Framework for Government.* Expresa sin mayor reparo que "*El control sistemático que la Tesorería General de la Seguridad Social realiza mediante la aplicación de los procedimientos habituales de gestión se ha complementado con la implantación de una metodología basada en el análisis del riesgo y con la incorporación de herramientas informáticas diseñadas de forma específica para el tratamiento del fraude*".

Por lo tanto, tenemos constancia de la elaboración de perfiles y patrones de comportamiento junto a análisis del riesgo en la comisión de posibles actuaciones fraudulentas. Veremos a continuación qué traducción puede tener la observación del Reglamento de IA y la Ley 15/2022.

3. ¿CÓMO VALORAMOS LA SITUACIÓN SI OBSERVAMOS EL REGLAMENTO DE IA Y LA LEY 15/2022?

Para abordar el asunto debemos partir de considerar que la modernización de la gestión de las prestaciones y de las técnicas de detección y persecución del error, del abuso o del fraude no se pueden llevar a cabo de cualquier manera y en todo caso, observando los marcos normativos europeos e internos.

[21] Disponible en: https://oecd-opsi.org/wp-content/uploads/2019/12/Documento-OCDE-finalV2.pdf

Comencemos con el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) nº 300/2008, (UE) nº 167/2013, (UE) nº 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial)[22], (en adelante, RIA).

Aunque los principios definidos como directrices éticas para una IA fiable por el Grupo independiente de expertos de alto nivel sobre IA creado por la Comisión no sean vinculantes en sentido estricto, inspiran, como no podía ser de otra manera, los pilares del RIA junto a la elaboración de los códigos de conducta que deberían respetarse a la hora del diseño y utilización de los sistemas: acción y supervisión humanas, solidez técnica y seguridad; gestión de la privacidad y de los datos; transparencia; diversidad, no discriminación y equidad; bienestar social y ambiental y rendición de cuentas.

Sentado lo anterior, valoremos si los sistemas que se están utilizando por parte del INSS para la gestión de las prestaciones de IT e IMV comportan un alto riesgo a tenor del RIA y por lo tanto quedarían sujetos a sus mandatos, que velan por evitar un uso opaco y descontrolado que descuide la depuración de los sesgos y pueda provocar efectos discriminatorios.

Si queremos alcanzar una interpretación teleológica resulta imprescindible leer detenidamente la prelación de considerandos del RIA. A efectos de lo que me interesa, uno de los más importantes es el considerando número 58, que ante el empleo de sistemas de IA fija como materia sobre la que hemos de prestar especial atención, entre otras, las prestaciones de Seguridad Social. Teniendo en cuenta la relación asimétrica entre la Administración Pública y las personas solicitantes de protección, el RIA se refiere a los últimos como personas que se encuentran en una posición de vulnerabilidad.

Por su importancia y novedad quiero destacar la literalidad de la siguiente afirmación incluida en el mismo considerando: *"La utilización de sistemas de IA para decidir si las autoridades deben conceder, denegar, reducir o revocar dichas prestaciones y servicios o reclamar su devolución, lo que incluye decidir, por ejemplo, si los beneficiarios tienen legítimamente derecho a dichas prestaciones y servicios, podría tener un efecto considerable en los medios de subsistencia de las personas y vulnerar sus derechos fundamentales, como el derecho a la protección social, a la no discriminación, a la dignidad humana o a la tutela judicial efectiva y, por lo tanto, deben clasificarse como de alto riesgo"*.

Si bien es cierto que acto seguido, en el mismo considerando 58º se advierte que el RIA no puede resultar un obstáculo para la innovación de la Administración Pública, también se afirma con rotundidad que ese objetivo de modernización no podrá efectuarse a costa del riesgo que supone para las personas, esto es, mediante sistemas que acarreen un alto riesgo para las personas.

Efectivamente, de forma general, emplear sistemas de IA para decidir sobre la concesión, denegación, reducción o revocaciones de prestaciones de la Seguridad Social puede generar nuevas formas de discriminación o discriminar sistemáticamente a determinadas personas y colectivos, resultando una manera de perpetuar los patrones discriminatorios históricos (por motivos como el origen racial o étnico, género, discapacidad, edad u orientación sexual).

En cuanto a la posible interpretación disyuntiva que se le puede dar al considerando 58º cuando finalmente menciona que al mismo tiempo debemos evitar obstaculizar la modernización de los sistemas, a mi modo de ver no deja margen de duda para la ponderación de intereses en la

[22] «DOUE» núm. 1689, de 12 de julio de 2024, páginas 1 a 144 (144 págs.)

que considero que prima notoriamente la protección ante la modernización de los sistemas de las administraciones.

En cualquier caso, comparto el planteamiento de Goerlich Peset[23], en cuanto a la calificación de alto riesgo como vía de equilibrio.

Planteada esta consideración 58ª del RIA, debemos atender al Anexo III del mismo cuerpo normativo, relativo a los sistemas de IA de alto riesgo al que nos remite el artículo 6.2 del mismo cuerpo normativo[24]. A su vez, entre el listado de sistemas de IA calificados como de alto riesgo, el apartado 5º se refiere al acceso a prestaciones públicas y concretamente, en su apartado a), a; "*sistemas de IA destinados a ser utilizados por las autoridades públicas o en su nombre para evaluar la admisibilidad de las personas físicas para beneficiarse de servicios y prestaciones esenciales de asistencia pública, incluidos los servicios de asistencia sanitaria, así como para conceder, reducir o retirar dichos servicios y prestaciones o reclamar su devolución".*

Parece que con lo expuesto puede alcanzarse la conclusión de que los sistemas de IA que viene empleando el INSS para la gestión de las prestaciones de IT e IMV comportan un alto riesgo conforme a las indicaciones del RIA y que deberían adaptarse a sus exigencias para tratar de evitar patrones sesgados con efectos discriminatorios.

¿Qué ocurre con los sistemas de IA que fueron incorporados en las entidades del sector público de forma previa a la entrada en vigor del RIA? El artículo 111 del RIA se refiere a dicho supuesto y no exime del cumplimiento de los requisitos que contiene, para lo que establece distintos plazos de adaptación, que dependen, entre otras cuestiones, del momento de su incorporación. En lo que aquí interesa, el apartado segundo *in fine* del mismo precepto establece que, en cualquier caso, los responsables del despliegue de los sistemas de IA de alto riesgo que vayan a destinarse a su uso por las autoridades públicas tendrán que adoptar las medidas que sean necesarias para cumplir con las obligaciones contenidas en el RIA antes del 2 de agosto de 2023.

Por el momento, el anteproyecto de Ley para el buen uso y la gobernanza de la inteligencia artificial no aborda este asunto ni contempla obligación alguna. Su exposición de motivos (concretamente, el párrafo segundo del apartado quinto *in fine*), refiere que el ámbito de aplicación de la ley incluye *"las personas jurídicas y entidades del sector público que actúen como operadores, según los define el propio Reglamento*" (se refiere al RIA). Si atendemos a los apartados cuarto y octavo del precepto tercero del RIA, a efectos del mismo hemos de entender como "responsable del despliegue" una persona física o jurídica, o autoridad pública, órgano u organismo que utilice un sistema de IA bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional y como "operador" a un proveedor, fabricante del producto, responsable del despliegue, representante autorizado, importador o distribuidor. Sin embargo, por parte del legislador español nada se menciona de las obligaciones a las que deberían quedar sujetas las Entidades Gestoras de las Seguridad Social si actúan como responsables del despliegue de sistemas de IA de alto riesgo (cuyo eje vertebrador es el artículo 26 del RIA).

En este punto procede interpretar la calificación de los sistemas de IA que se utilizan para la detección de presuntas actuaciones fraudulentas.

[23] Goerlich Peset, J.M.: "Reglamento de inteligencia artificial e intervención pública en las relaciones laborales". *LABOS Revista De Derecho Del Trabajo Y Protección Social*, vol. 5, 2024. Pp. 228-242.

[24] "*Además de los sistemas de IA de alto riesgo a que se refiere el apartado 1, también se considerarán de alto riesgo los sistemas de IA contemplados en el anexo III.*"

Se considera que el legislador europeo quizás no ha sido suficientemente preciso a la hora de definir qué se entiende por puntuación ciudadana. La necesidad de protección ante el supuesto de persecuciones inspectoras injustificadas basadas en elaboraciones de perfiles discriminatorios es obvia, pero, ¿podría extenderse esta consideración a la perspectiva que estamos tratando?

El artículo 3.52) del RIA nos remite a la definición de "elaboración de perfiles" contenida en el artículo 4.4) del Reglamento General de Protección de Datos[25], que incluye cuestiones directamente relacionadas con el acceso y la gestión de las dos prestaciones que vengo poniendo como ejemplo: tratamiento automatizado de datos personales para realizar predicciones sobre, entre otras, salud y situación económica[26].

Por su parte, el considerando núm. 31 del RIA advierte de los tratos discriminatorios que pueden resultar de los sistemas de IA que posibilitan que los agentes (a efectos de lo que aquí interesa, públicos), manejen puntuaciones que califiquen a las personas físicas. Tales modelos evalúan o clasifican a las personas físicas sobre sus comportamientos o características personales (sean conocidas, inferidas o predichas). El resultado puede servir de motivación para que de facto reciban un trato desfavorable. En definitiva, estas prácticas pueden vulnerar el derecho a la dignidad y a la no discriminación, así como los valores de igualdad y justicia. Consecuentemente, mediante el considerando citado se expresa que los sistemas de IA que sirvan para llevar a cabo dichas prácticas de puntuación inaceptables que provocan situaciones perjudiciales deben prohibirse (quedando prohibido a tenor del artículo 5.1 c)[27] y d)[28] del RIA).

Las propias explicaciones de la TGSS a las que se ha hecho referencia con anterioridad mencionan expresamente la definición de perfiles, elaboración de patrones de comportamiento y atribución de riesgos para perseguir el fraude mediante la herramienta de IA denominada *SAS Fraud Framework*, lo que yo encuadraría en el ámbito de la prohibición delimitado por el artículo quinto del RIA. Aunque el citado apartado d) se refiera a la prohibición del empleo de sistema de IA para realizar evaluaciones de riesgos de personas físicas con el fin de valorar o predecir el riesgo de que una persona física cometa un delito (y no una infracción administrativa) basándose únicamente en la elaboración del perfil de una persona física en la evaluación de los rasgos y características de su personalidad, lo razonable sería darle una interpretación amplia para poder aplicarlo de forma

[25] Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos y por el que se deroga la Directiva 95/46/CE. «DOUE» núm. 119, de 4 de mayo de 2016, páginas 1 a 88 (88 págs.)

[26] "*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*"

[27] "*c) la introducción en el mercado, la puesta en servicio o la utilización de sistemas de IA para evaluar o clasificar a personas físicas o a colectivos de personas durante un período determinado de tiempo atendiendo a su comportamiento social o a características personales o de su personalidad conocidas, inferidas o predichas, de forma que la puntuación ciudadana resultante provoque una o varias de las situaciones siguientes: i) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos de personas en contextos sociales que no guarden relación con los contextos donde se generaron o recabaron los datos originalmente, ii) un trato perjudicial o desfavorable hacia determinadas personas físicas o colectivos de personas que sea injustificado o desproporcionado con respecto a su comportamiento social o la gravedad de este;"*

[28] "*d) la introducción en el mercado, la puesta en servicio para este fin específico o el uso de un sistema de IA para realizar evaluaciones de riesgos de personas físicas con el fin de valorar o predecir el riesgo de que una persona física cometa un delito basándose únicamente en la elaboración del perfil de una persona física o en la evaluación de los rasgos y características de su personalidad; esta prohibición no se aplicará a los sistemas de IA utilizados para apoyar la valoración humana de la implicación de una persona en una actividad delictiva que ya se base en hechos objetivos y verificables directamente relacionados con una actividad delictiva; …"*

extensiva a toda elaboración de perfiles que persiga discriminatoriamente a las personas, sea en vía administrativa o penal.

Si no se comparten estas dos interpretaciones sobre los sistemas de IA de alto riesgo y los sistemas susceptibles de prohibición, en todo caso, a nivel interno es menester tener presente de forma transversal el contenido de la Ley 15/2022, de 12 de julio, integral para la igualdad de trato y la no discriminación que resulta de aplicación a las Entidades Gestoras tanto desde el prisma subjetivo como objetivo.

Veamos sus traducciones. Dedica su artículo 23º a la IA y los mecanismos de toma de decisión automatizados, cuyo contenido es muy positivo y cuya aplicación hasta el momento parece inexistente. Se refiere a que las Administraciones Públicas "favorecerán" mecanismos para que los sistemas algorítmicos involucrados en su toma de decisiones observen criterios de minimización de sesgos, transparencia y rendición de cuentas. El espectro al que se refiere es completo porque establece que dichos mecanismos tratarán desde el diseño de los algoritmos hasta los datos de entrenamiento observando su potencial impacto discriminatorio (para lo que se indica que *"se promoverá la realización de evaluaciones de impacto que determinen el posible sesgo discriminatorio"*). También contiene expresamente el deber de promover el uso de una IA ética, confiable y respetuosa con los Derechos Fundamentales en la línea de las recomendaciones de la UE. Pero lo más llamativo es el segundo apartado, que fija un mandato expreso sobre el deber de priorizar la transparencia en todas las fases (diseño, implementación y evaluación) cuando precisamente y como viene insistiendo la doctrina mencionada *ut supra*, existe una opacidad preocupante sobre los sistemas que utilizan las Entidades Gestoras de la Seguridad Social. Llama la atención el incumplimiento de todas estas previsiones protectoras acaecido hasta el momento.

Sentado lo anterior, además de observar las obligaciones fijadas en los dos niveles predichos considero que una de las líneas fundamentales de actuación debería consistir en realizar estudios de impacto de género de forma que podamos evitar los efectos discriminatorios estructurales en la línea de lo que evidencian los estudios científicos referenciados, que afectan a la salud y el bienestar de las mujeres. Precisamente, así lo prevé el artículo 4.4 de la Ley 15/2022[29].

Por otro lado, opino que no deberíamos exigir (aunque *ex* artículo 30.1º de la Ley 15/2022[30] es a lo que debemos atenernos por el momento) que la persona afectada, o más fácilmente, el interesado, tenga que aportar indicios fundados para que se produzca la inversión de la carga de la prueba cuando estamos hablando de posibles mecanismos complejos de discriminación indirecta e interseccional en una relación tan asimétrica como la existente entre los solicitantes de prestaciones como la IT o el IMV y el INSS, la TGSS o la ITSS. Si se diera el supuesto de inversión de la carga de la prueba, en los casos que vengo poniendo de supuestos, por ejemplo, el INSS, debería acreditar que el sistema no cae en errores o sesgos con efectos discriminatorios.

Junto a los preceptos citados, mientras no se aborde responsablemente la cuestión, también resultará imposible garantizar el cumplimiento del contenido del artículo 25 de la Ley 15/2022 relativo a las medidas de protección y reparación frente a la discriminación. ¿Cómo vamos a hablar de detección y corrección de patrones discriminatorios si no hablamos de las herramientas que se están

[29] "*En las políticas contra la discriminación se tendrá en cuenta la perspectiva de género y se prestará especial atención a su impacto en las mujeres y las niñas como obstáculo al acceso a derechos como la educación, el empleo, la salud, el acceso a la justicia y el derecho a una vida libre de violencias, entre otros*".

[30] "*De acuerdo con lo previsto en las leyes procesales y reguladoras de los procedimientos administrativos, cuando la parte actora o el interesado alegue discriminación y aporte indicios fundados sobre su existencia, corresponderá a la parte demandada o a quien se impute la situación discriminatoria la aportación de una justificación objetiva y razonable, suficientemente probada, de las medidas adoptadas y de su proporcionalidad"*.

utilizando? Por no hablar del régimen de responsabilidad al que se hace referencia y el resarcimiento del daño en un problema que puede ser estructural y sistémico.

La debida transparencia exige que los sistemas de IA posibiliten una trazabilidad en los términos del artículo 12 del RIA en relación con el artículo 19 del mismo texto y una "explicabilidad" desde el diseño del sistema, pasando por su desarrollo y la gestión del responsable del despliegue hasta las personas afectadas por la aplicación del sistema y la ciudadanía en general. Recordemos que la falta de información se puede interpretar como una discriminación por inacción *ex* artículo 4.1 de la Ley 15/2022.

Parte de la doctrina expresa acertadamente que los algoritmos no deben tener en cuenta únicamente razonamientos matemáticos, de programación o de estadística, ya que deben observar los de la Sociología, la Ciencia Política, la Filosofía y el Derecho[31]. Partiendo de esta premisa, llama la atención que más adelante en dicho estudio se plantee que es necesario que concurra una correlación entre lo que se denomina *fairness* artificial y la *fairness* legal. Sostiene que para sustentar que concurre discriminación necesariamente hay que establecer el nexo causal entre la intencionalidad y el resultado y añade, que encontrar estos elementos del *fairness* en el *Machine Learning* es complicado. Pues bien, darle este enfoque restrictivo al problema no parece lo más acertado. Considero que no deberíamos exigir la acreditación de la intencionalidad para dar por ciertas las dinámicas sesgadas y discriminatorias que venimos describiendo y que se agravan con el uso descontrolado de la analítica avanzada. De esa forma dejaríamos fuera probablemente todo lo relativo a la discriminación indirecta o la que responde a errores de muestreo o de programación. También cabe matizar que exigir transparencia y búsqueda de la forma más adecuada de utilizar estas herramientas no implica tener aversión a los algoritmos. La aversión es hacia la opacidad que hasta el momento rige sobre el asunto, lo que es especialmente grave tratándose de la Administración Pública.

Cotino Hueso[32] nos está aportando importantes razonamientos en materia de la utilización de sistemas de IA en el sector público en general y me interesa trasladar varios de ellos al ámbito de la Seguridad Social.

Resulta interesante el planteamiento de la adopción de medidas positivas o que podríamos llamar de discriminación positiva y que podríamos ejemplificar, al menos, con la intención que tuvo la medida, con la búsqueda de las familias potencialmente cumplidoras de los requisitos de acceso al IMV. No obstante, como he expuesto anteriormente, las dificultades acaecidas desde el inicio de la gestión de la prestación para que la ayuda llegue a las familias provocan que si el rastreo y valoración del cumplimiento de los requisitos depende de un algoritmo que se retroalimenta de los resultados que obtiene, nos alejemos todavía más de los supuestos susceptibles de protección e ignorados por el sistema.

Por dicho motivo, deberíamos conocer las fuentes de la información, los criterios de valoración y la relación con el resultado en el proceso técnico para la toma de decisiones o recomendaciones. Todo ello, contando con que el sistema motive la aplicación de los criterios y justifique los pasos que da.

[31] Belloso Martín, N.: "Sobre Fairness y Machine Learning: El Algoritmo ¿Puede (y Debe) Ser Justo?". *Anales de la Cátedra Francisco Suárez*, Nº 57, 2023. Pp. 7-38.

[32] Cotino Hueso, L.: "Discriminación, sesgos e igualdad de la inteligencia artificial en el sector público" en *Inteligencia artificial y sector público: retos, límites y medios*. Valencia, Tirant lo Blanch, 2023. Pp. 257-351.

Como bien defiende el citado autor, más vale prevenir que discriminar. En este sentido, no puede ser más acertada la visión que aporta a lo largo de su detallado análisis sobre el problema multifactorial al que nos enfrentamos.

Concretamente, como elemento central quiero destacar la idea de que ciertos requisitos para la exigibilidad de un esfuerzo normativo que provea garantías ante el uso de estos sistemas por la Administración Pública suponen simple y llanamente una huida del Derecho.

Ejemplo de ello es el nivel de intervención humana y la distinción del grado de participación y por lo tanto de autonomía del sistema. Llevado a los ejemplos en los que me he centrado, las prestaciones de IT y de IMV, parece que, con lo que sabemos, en ambos casos hablamos de una sistematización parcial o automatización instrumental. En todo caso, los sistemas de IA que se empleen directa o indirectamente para la toma de decisiones en materia prestacional deberían ser evaluados para velar por el respeto a los Derechos Fundamentales.

En este sentido, el concepto denominado "sesgo de automatización" recogido por el artículo 14.4.b) del RIA es esencial, ya que no debemos confiar excesivamente o de forma automática en los resultados del sistema de IA de alto riesgo especialmente si dicho resultado se utilizará como apoyo para la toma de decisiones (aunque la decisión final se tome por un ser humano). Ligado a la importancia del factor humano, controlar la sensibilidad y la consciencia de quien entrena el algoritmo es esencial tanto si se trata de un servicio interno como de uno externalizado y no se puede dejar exclusivamente en manos de los programadores (que, aunque fuese deseable, en principio no cuentan con el desarrollo de competencias sobre estas materias).

En definitiva, lo cierto es que no tenemos certeza sobre el nivel de intervención humana en el empleo de estas herramientas por las Entidades Gestoras debido a la opacidad de las prácticas. En todo caso, me pregunto si vamos a condicionar el control y una regulación decente de estas prácticas al nivel de intervención humana o si basta con que valoremos de una vez por todas el impacto de la automatización de los sistemas de evaluación, concesión, revocación de prestaciones y servicios esenciales de asistencia pública en la salud de las personas y la subsistencia de las familias (especialmente en ejemplos como los que he traído a colación). Estimo que lo razonable es lo segundo y que no podemos dejar que prosperen las dinámicas discriminatorias y acarreen consecuencias desastrosas e impropias de la filosofía de nuestro sistema. Este razonamiento es aplicable tanto al reconocimiento y gestión de las prestaciones como a la priorización de los objetivos para la aplicación de la Ley.

Ante lo que todo lo que aquí se ha expuesto podemos ir más allá. En el supuesto del empleo de IA predictiva por parte de la TGSS, la ITSS o la Inspección Tributaria, ¿vamos a permitir que se impulsen inspecciones por las que se termine persiguiendo a determinados colectivos por el perfil del que parten los sistemas nutridos con información sesgada que no tiene en cuenta los problemas estructurales de la sociedad como ha ocurrido en otros países? Como se ha podido observar, el RIA es bastante claro en cuanto a los supuestos de prohibición de la elaboración de perfiles y atribución del riesgo, pero viendo la huida del Derecho que puede suponer, como bien plantea el autor citado *ut supra*, la discusión en torno al rol del ser humano en el empleo por parte del sector público de estas herramientas en cuestiones tan delicadas, lo planteo. ¿Vamos a permitir que resulte una herramienta por la que se perpetúen y multipliquen los patrones discriminatorios? Considero que los planes estratégicos de la ITSS deberían prever mecanismos de control para evitar la perpetuación de los patrones discriminatorios más allá de las vagas referencias a la IA.

Además, deberíamos ocuparnos de valorar el peligro en contextos de tratamiento masivo de datos y toma de decisiones esenciales para el bienestar social. Como bien explica el considerando 110º del RIA, cuanto mayor sea la capacidad del modelo mayores serán los riesgos sistémicos que pueden materializarse en cualquier momento del ciclo y que quedan determinados por su diseño, su uso y su alcance.

A su vez, convendría prestar atención a todas las fases de implementación y uso del sistema sin esperar a que los efectos sean irreparables (partiendo del diseño del sistema público o la contratación del mismo si se externaliza y considerando y evitando posibles sesgos en los términos del precepto noveno del RIA[33]).

La analítica de datos tiene que tener en cuenta la sociología y los problemas estructurales y para ello debe contar con equipos multidisciplinares que se ocupen de respetar la diversidad, la no discriminación y la equidad. Especialmente si nos referimos al desarrollo de sistemas en la Seguridad Social, tiene que observarse la igualdad de acceso, la igualdad de género y la diversidad casuística. En la misma línea, en cuanto a la calidad de los datos pertinentes (de entrenamiento, validación y prueba) y las fuentes de información, debe garantizarse una mirada inclusiva que refleje la diversidad de los colectivos. Para ello necesitamos mecanismos que permitan detectar la discriminación (especialmente la indirecta por ser la más invisibilizada). Como bien lo contextualiza el considerando 67 del RIA, los datos constituyen los cimientos de estos sistemas de entrenamiento de modelos de IA. Es esencial evitar dejaciones en este sentido porque provocarán una fuente de discriminación. El mismo considerando añade que los conjuntos de datos deben tener en cuenta el contexto en el que se prevé aplicar la herramienta, las características o elementos particulares del entorno geográfico, contextual, conductual o funcional específico. También quiero destacar el artículo 10.2.f) del RIA, que se refiere a la necesidad de atender a los sesgos que puedan afectar a la salud y la seguridad de las personas y sean discriminadas quedando vulnerados los Derechos Fundamentales. Precisamente, el contexto de aplicación al que me he referido por medio del presente análisis es extremadamente delicado y requiere especial cautela y dedicación.

En cuanto al desarrollo de la aplicación del modelo, deberíamos dejar de asumir la opacidad de los algoritmos y exigir que los sistemas funcionen conforme a parámetros transparentes. De igual forma, pensar en cómo utilizar los algoritmos para la detección de los patrones discriminatorios. De esta forma, la evaluación *ex post* también resulta esencial porque los resultados reflejarán las necesidades de corregir los sistemas y es evidente que para ello la intervención humana es imprescindible. La transparencia también debe reflejarse en esta fase y especialmente, en supuestos como los planteados en los que las personas interesadas afectadas por el resultado, como ya hemos visto, podrán alegar indicios de una vulneración. Es esencial atender al contenido del artículo 15.4, párrafo 3º del RIA[34] en clave social, como se ha intentado expresar por medio de este estudio.

Por todo ello deben establecerse mecanismos de control de los sistemas que se utilizan por parte de las Entidades Gestoras de la Seguridad Social y la ITSS e intentar acabar con los sistemas opacos en estos contextos esenciales que impactan en el derecho a la seguridad social.

[33] Los desarrolladores, desde su posición de proveedores de los sistemas deberán facilitar documentación técnica a la Administración Pública como responsable del despliegue *ex* arts. 11 y 3 del RIA.

[34] "*Los sistemas de IA de alto riesgo que continúan aprendiendo tras su introducción en el mercado o puesta en servicio se desarrollarán de tal modo que se elimine o reduzca lo máximo posible el riesgo de que los resultados de salida que pueden estar sesgados influyan en la información de entrada de futuras operaciones (bucles de retroalimentación) y se garantice que dichos bucles se subsanen debidamente con las medidas de reducción de riesgos adecuadas."*

Sentadas estas consideraciones, quedamos a expensas de las actuaciones que se lleven a cabo para el cumplimiento del marco normativo sobre la materia (especialmente en lo referido a las obligaciones que a tenor del artículo 26 del RIA tendrán las Entidades Gestoras de la Seguridad Social que utilicen estos sistemas por ocupar la posición de responsables del despliegue de los sistemas de IA de alto riesgo). Sabemos que se ha puesto en marcha la Agencia Española de Supervisión de Inteligencia Artificial (AESIA) con el objeto de cumplir con las exigencias del artículo 59 del RIA. Tendremos que esperar para valorar su nivel de actividad y eficacia. A su vez, veremos el papel que juega la Autoridad Independiente para la Igualdad de Trato y la no Discriminación y si favorece la investigación de supuestos como los que se han tratado, que hasta ahora han pasado más o menos inadvertidos.

4. BIBLIOGRAFÍA

Amnistía Internacional: *Injusticia codificada. vigilancia y discriminación en el estado de bienestar automatizado de Dinamarca*. Londres, Amnesty International Ltd., 2024.

Belloso Martín, N.: "Sobre Fairness y Machine Learning: El Algoritmo ¿Puede (y Debe) Ser Justo?". *Anales de la Cátedra Francisco Suárez*, Nº 57, 2023. Pp. 7-38.

Cabanillas-Montferrer, T., et al.: "El sesgo de género en la asistencia sanitaria: definición, causas y consecuencias en los pacientes". *MUSAS. Revista de Investigación En Mujer, Salud y Sociedad*, 2022, vol. 7, nº 1, pp. 106-129.

Cotino Hueso, L.: "Discriminación, sesgos e igualdad de la inteligencia artificial en el sector público" en *Inteligencia artificial y sector público: retos, límites y medios*. Valencia, Tirant lo Blanch, 2023. Pp. 257-351.

De Miguel Beriain, I.: "Medicina personalizada, algoritmos predictivos y utilización de sistemas de decisión automatizados en asistencia sanitaria: Problemas éticos". *Dilemata*, 2019, nº 30, pp. 93-109.

Escudero Rivas, C.: "El análisis predictivo en las relaciones laborales y de seguridad social" en VV.AA.: *VII Congreso Internacional y XX Congreso Nacional de la Asociación Española de Salud y Seguridad Social. Las transformaciones de la Seguridad Social ante los retos de la era digital,* Murcia, Laborum, 2023. Tomo I, pp. 29-41.

Fernández orrico, F.J.: "La IA como instrumento de predicción en materia de altas médicas. Un análisis a la vista de la normativa sobre protección de datos y la propuesta de Reglamento (UE) de IA" en: VV.AA.: *VII Congreso Internacional y XX Congreso Nacional de la Asociación Española de Salud y Seguridad Social. Las transformaciones de la Seguridad Social ante los retos de la era digital*, Murcia, Laborum, 2023. Tomo I, pp. 133-153.

Fernández Ramírez, M.: "Inteligencia Artificial, algoritmos predictivos y gestión tecnológica de la Seguridad Social" en *Las transformaciones de la Seguridad Social ante los retos de la era digital: VII Congreso Internacional y XX Congreso Nacional de la Asociación Española de Salud y Seguridad Social*. Murcia, Laborum, 2023. Tomo I, pp. 155-174.

Goerlich Peset, J.M.: "Reglamento de inteligencia artificial e intervención pública en las relaciones laborales". *LABOS Revista De Derecho Del Trabajo Y Protección Social*, vol. 5, 2024. Pp. 228-242.

Igartua Miró, M.T.: "El enfoque de género y la enfermedad profesional: reflexiones desde el ordenamiento español". *Revista Internacional y Comparada de Relaciones Laborales y Derecho del Empleo*, ADAPT University Press, vol., 1, nº2, 2023.

Martín López, J.: "Inteligencia artificial, sesgos y no discriminación en el ámbito de la inspección tributaria". *Crónica Tributaria*, nº 182, 2022. Pp. 51-89.

Mesa, S.: *Silencio administrativo. La pobreza en el laberinto burocrático*. Barcelona, Editorial Anagrama, S.A.U., 8ª ed., 2019.

Pardo García, J.: "La analítica avanzada de datos en la Seguridad Social"*, Astic*, 2018, disponible en: https://www.astic.es/wp-content/uploads/2018/06/boletic82-monografico4-juanpardo.pdf

Pérez-Ugena Coromina, M.: "Sesgo de género (en IA)". *EUNOMÍA. Revista en Cultura de la Legalidad*, 2024, nº 26, pp. 311-330.

Romero Coronado, J.: "Impacto del Big Data y la Inteligencia Artificial en la gestión de la Seguridad Social". *Revista de Derecho de la Seguridad Social, Laborum*. Extraordinaria 6, 2024, pp. 49-70.

Tan, E., et al. "Artificial intelligence and algorithmic decisions in fraud detection: An interpretive structural model". *Data & policy*, *Cambridge University Press*, 2023, vol. 5, e25.

Ünal, C., Erbuğa, G.S.: "Detection and Prevention of Medical Fraud using Machine Learning". *Acta Infologica*, 2024, vol. 8, nº 2, p. 100-117.

Villar Cañada, I.M.: "El impacto tecnológico en el ámbito sociolaboral: ¿obstáculo u oportunidad para la sostenibilidad del sistema de pensiones?". *Revista de Derecho de la Seguridad Social, Laborum*. Extraordinaria nº 7 (2024): 241-261.

Villar Cañada, I.M.: "La digitalización y los sistemas de protección social: oportunidades y desafíos". *Revista de Trabajo y Seguridad Social. CEF*, 459, 173-205.