LIBRO BLANCO

+IGUAL

UN ENFOQUE PRÁCTICO, ÉTICO Y NORMATIVO PARA DESARROLLAR UN ESTÁNDAR DE CERTIFICACIÓN DE IA EN EL ÁMBITO LABORAL

RESUMEN EJECUTIVO

Julio 2025

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE DERECHOS SOCIALES, CONSUMO Y AGENDA 2030
Plan de Recuperación, Transformación y Resiliencia
Comunidad de Madrid
CONSEJERÍA DE FAMILIA, JUVENTUD Y ASUNTOS SOCIALES

El proyecto IA+Igual ha realizado un análisis empírico del impacto de los sesgos, conscientes e inconscientes, en el funcionamiento ético de las herramientas de IA aplicadas a procesos en el ámbito del mercado laboral. Con los resultados del análisis se han definido estrategias que permiten promover los derechos de equidad e inclusión, la transformación digital ética y la inserción sociolaboral de todos los colectivos vulnerables y en riesgo de exclusión.

El proyecto ha sido financiado con fondos Next Generation, a través del área de Innovación Social de la Comunidad de Madrid, y del Plan de Resiliencia y Recuperación del Gobierno de España, y se ha desarrollado entre julio de 2023 y junio de 2025.

Uno de los objetivos del proyecto era la elaboración de un Libro Blanco que analiza la relación entre la evolución humana, los sesgos cognitivos y el impacto de la inteligencia artificial en el ámbito del mercado laboral y los Recursos Humanos (RR. HH.). Titulado *Un enfoque práctico, ético y normativo para desarrollar un estándar de certificación de la IA en el ámbito de laboral*, y coordinado por el Instituto de Ciencia de los Datos y la Inteligencia Artificial (DATAI) de la Universidad de Navarra, este Libro Blanco se ha desarrollado con una metodología rigurosa y participativa. Además, se ha utilizado un enfoque multidisciplinar y herramientas de IA que aseguran la calidad del contenido.

Este documento, de alto valor en el ámbito de la innovación social, sienta las bases para el desarrollo de un estándar de certificación ética específica para herramientas de IA aplicadas a procesos del mercado laboral y los RR. HH; además de proponer líneas de actuación tanto para la empresa privada, como las administraciones públicas, que permitan avanzar en el cumplimiento de los ODS y sus objetivos de inclusión, equidad y diversidad.

El documento que tiene en sus manos es un Resumen Ejecutivo de su contenido. Le emplazamos a consultarlo y enriquecerse con sus aportaciones.

Nota editorial: cada capítulo del Libro Blanco tiene entidad propia lo que facilita la selección o el orden de lectura.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España
Ministerio de Derechos Sociales, Consumo y Agenda 2030
Plan de Recuperación, Transformación y Resiliencia
Comunidad de Madrid
Dirección General de Evaluación, Calidad e Innovación
Consejería de Familia, Juventud y Asuntos Sociales

ÍNDICE

RESUMEN EJECUTIVO

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

INTRODUCCIÓN

La forma en que vivimos hoy es el resultado de millones de cambios evolutivos y decisiones que han moldeado nuestra identidad. Comprender esta realidad es clave para ejercer el control sobre nosotros mismos y nuestro entorno.

Nuestro cerebro está programado para sobrevivir y simplificar la información abrumadora que recibe a través de los sentidos, para detectar peligros. De esta forma, los sesgos cognitivos se convierten en atajos mentales que influyen en nuestra percepción y toma de decisiones. En 1972, Kahneman y Tverskyen los definieron como interpretaciones erróneas sistemáticas de la realidad. Influenciados por la herencia genética y el aprendizaje, los sesgos alimentan la discriminación y limitan la visión en la investigación y el desarrollo tecnológico.

Una vez que entra en escena la IA generativa (IAg), los sesgos se incorporan de forma consciente o inconsciente a los modelos de entrenamiento. De esta forma, contribuyen a perpetuar estereotipos y desigualdades que en ámbito del mercado laboral y los procesos de RR. HH., han sido calificados como de Alto Riesgo por la Unión Europea.

En efecto, la irrupción de la IAg ha planteado retos éticos y sociales en todas las economías del mundo a los que se da respuesta, principalmente, desde tres perspectivas:

- Control socioeconómico: La IA se utiliza para supervisar y clasificar personas, perpetuando sistemas políticos y económicos.
- Control sociopolítico: La falta de reglas y normas éticas en el uso de la IA genera riesgos significativos.
- La persona, en el centro del algoritmo: La Unión Europea busca regular la IA para garantizar su uso ético y el respeto a los derechos humanos.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

1.- IA EN EL ÁMBITO LABORAL

La Inteligencia Artificial (IA) está revolucionando el mercado laboral. La gestión de Recursos Humanos está pasando de un enfoque administrativo a uno predictivo y estratégico. Esto permite a las organizaciones, por ejemplo, optimizar procesos, facilitar la identificación de potenciales empleados y la optimización de procesos de selección; el desarrollo de modelos personalizados de aprendizaje y desarrollo profesional, y generar entornos de trabajo más adaptativos e inclusivos.

Antes de la IA, la gestión de Recursos Humanos se basaba en enfoques manuales y subjetivos, basados en la experiencia directa. La informatización y el uso de datos comenzaron a cambiar esta dinámica, primero hacia un enfoque más analítico con sistemas de gestión que priorizaban la eficiencia operativa; y luego hacia un enfoque más analítico, con la irrupción de la cultura del dato.

Ahora, con la irrupción de la IAg, se produce una nueva transformación de la gestión del talento que impacta, también, en la toma de decisiones. Sin embargo, esta tecnología avanza tan rápido, de forma tan disruptiva y transversal, que es difícil contar con el conocimiento, -y mucho menos la experiencia- para implantarla con seguridad. La moda, las tendencias y la promesa de una mejora en la eficiencia y productividad empuja a la mayoría de las organizaciones a tomar decisiones de compra e implantación de herramientas tecnológicas sin entender bien el impacto que puede tener dicha herramienta en la organización.

Es crucial una toma de decisiones informada, que tiene como punto de partida entender qué es y qué no es IA; porqué la IA no es infalible; dónde y con qué datos de entrenamiento se ha desarrollado la herramienta; y de qué manera los sesgos pueden haberse introducido de forma consciente o inconsciente en el modelo algorítmico, generando una discriminación estructural. Desconocer esta información básica, además de suponer un posible incumplimiento normativo, implica sanciones económicas elevadas y un riesgo reputacional para la organización.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

1.1.Aplicaciones de IA en la Gestión del Talento

En el ámbito laboral, las herramientas soportadas en IA están transformando la gestión del talento a lo largo de todo el ciclo de vida profesional, desde la atracción y selección de personas hasta la desvinculación. La IA permite:

- Personalizar itinerarios formativos y anticipar necesidades de desarrollo.
- Identificar patrones de desempeño y predecir el potencial.
- Ayudar a identificar patrones de malestar y desconexión emocional.
- Personalizar la comunicación, mejorando la eficacia de los mensajes.
- Optimizar procesos como la gestión de turnos y el control horario.
- Anticipar situaciones de riesgo y generar recomendaciones personalizadas
- Agilizar el análisis de convenios y simular escenarios de negociación

La automatización de los procesos relacionados con la gestión de talento conlleva riesgos que hacen necesario implementar mecanismos de control y supervisión humana, y de explicabilidad algorítmica.

A continuación detallamos los potenciales beneficios Vs los retos de la aplicación de Inteligencia Artificial a diversos procesos de gestión del talento.

Riesgos y beneficios de IA aplicada a procesos de RR. HH.

Fase del ciclo de vida	Aplicación de IA	Beneficios potenciales	Riesgos y alertas
Atracción del talento	Matching algorítmico de CV, análisis predictivo de desempeño	Eficiencia en el filtrado; ampliación de perfiles afines	Exclusión de trayectorias atípicas; sesgos en datos históricos
Onboarding	Personalización de itinerarios, bots de bienvenida, IA conversacional	Integración rápida y adaptada a cada rol	Experiencias impersonales; desajuste con expectativas culturales

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Desarrollo y carrera	Recomendación de itinerarios profesionales, análisis de potencial, evaluación automatizada del desempeño	Visibilización de talento interno; proyección de carrera; alineación entre desempeño y oportunidades	Reproducción de techos de cristal; penalización de carreras no lineales; sesgos en la evaluación
Itinerarios formativos	Recomendación personalizada de contenidos, sistemas de tutoría virtual adaptativa, análisis automatizado de progreso	Formación alineada con el perfil y ritmo del empleado; identificación temprana de necesidades de aprendizaje	Dificultad en evaluar la transferencia real al puesto de trabajo; no visibilizar necesidades no previstas en el modelo; sesgos en el seguimiento del progreso
Productividad	Monitorización de tareas, análisis de eficiencia por procesos, predicción de cuellos de botella	Optimización de recursos; identificación de mejoras operativas; toma de decisiones basada en datos	Reducción del trabajo a métricas cuantitativas; vigilancia excesiva; presión constante sin considerar el contexto
Absentismo	Detección de patrones de ausencias, análisis predictivo de factores de riesgo, correlación con condiciones de trabajo	Prevención proactiva; identificación de causas organizativas; diseño de intervenciones tempranas	Estigmatización de perfiles; decisiones preventivas injustificadas; confusión entre causas personales y estructurales
Planes de sucesión	Predicción de relevo generacional, análisis de liderazgo emergente	Anticipación y continuidad organizativa	Favorecimiento de perfiles homogéneos; opacidad en criterios
Rotación y offboarding	Predicción de rotación no deseada, análisis de causas de salida, modelado de abandono, retroalimentación automatizada	Mejora de retención; rediseño organizativo; anticipación de problemáticas estructurales	Uso invasivo de datos sensibles; decisiones preventivas injustificadas; retroalimentación mal interpretada o descontextualizada

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE DERECHOS SOCIALES, CONSUMO Y AGENDA 2030
Plan de Recuperación, Transformación y Resiliencia
Comunidad de Madrid
Dirección General de Evaluación, Calidad e Innovación
CONSEJERÍA DE FAMILIA, JUVENTUD Y ASUNTOS SOCIALES

1.2. Gobernanza del dato en IA para RR. HH

La Gobernanza del dato es una de las claves para asegurar la efectividad y explicabilidad de la inteligencia artificial en RR. HH., puesto que garantiza la trazabilidad y supervisión de los datos utilizados. Para ello, se requiere una estructura técnica sólida que cumpla con la normativa y reproduzca las mejores prácticas.

Un modelo de Gobernanza del dato es un conjunto de principios, roles y procesos que aseguran la calidad y protección de los datos en una organización. Contar con un modelo de Gobernanza del dato supone responder a las exigencias del Reglamento Europeo de Inteligencia Artificial y el GDPR que establecen la obligación de asegurar la calidad y trazabilidad de los datos en sistemas de IA de alto riesgo, como es el caso de su uso en los procesos de RR. HH.

Estos modelos, tienen que cumplir al menos con estas líneas de actuación:

- Definir dos roles técnicos fundamentales para la gestión y calidad de los datos: Data Owners y Data Stewards.
- Definir e implantar mecanismos que permitan controlar de manera exhaustiva el ciclo de vida del dato, desde su captura hasta su almacenamiento.
- Definir e implantar mecanismos de trazabilidad y supervisión, como dashboards y auditorías para garantizar la equidad y explicabilidad de los algoritmos.

1.3 El impacto de la IA en materia de diversidad e inclusión

Las estrategias de Responsabilidad Corporativa, los criterios ESG (Environmental, Social, and Governance) y la alineación con los Objetivos de Desarrollo Sostenible de Naciones Unidas (ODS) generan un marco de compromiso con la igualdad, la diversidad y la inclusión sociolaboral que cada

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

vez más empresas, incorporan a estrategias corporativas y recogen en sus Memorias de Estados de Información no Financiera (EINF).

La aplicación de herramientas de IA en el ámbito laboral supone un nuevo reto en la consecución de estos objetivos puesto que, si no se implementan las salvaguardas adecuadas, se pueden perpetuar y agravar las desigualdades sociolaborales existentes.

La falta de representatividad de colectivos desfavorecidos o en riesgo de exclusión en los datos de entrenamiento de la IA, puede llevar a decisiones automatizadas que excluyan a estas personas que tienen trayectorias profesionales no normativas. Los datos históricos, especialmente en el ámbito laboral, están sesgados por el hecho objetivo del respeto a la normativa relacionada con la Protección de Datos. Estos sesgos penalizan los perfiles profesionales atípicos, como mujeres con trayectorias interrumpidas o personas mayores sin experiencia digital. Estos ejemplos pueden extrapolarse a otros colectivos desfavorecidos, como exreclusos, personas con enfermedades crónicas, etc.

La realidad es que la normativa de protección de datos, definida en un entorno analógico previo al actual, limita la inclusión de estos colectivos vulnerables a los conjuntos de entrenamiento, perpetuando así su invisibilidad.

Tomar conciencia y entender por qué se produce este sesgo, es el primer paso para adoptar un enfoque crítico hacia la representatividad en los datos de IA y mitigar el impacto indeseado de los sesgos.

En el transcurso de la investigación empírica desarrollada en el marco del Proyecto IA+Igual, se han constatado varias estrategias de actuación que dan respuesta a este reto:

- La validación diferenciada que permite evaluar el rendimiento de los algoritmos en subgrupos poblacionales.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

- Revisión de los datos de entrenamiento desde una perspectiva crítica para garantizar la equidad y desarrollar mecanismos de validación que evalúen el rendimiento algorítmico por subgrupos.
- Creación de conjuntos de validación específicos que ayuden a detectar sesgos asimétricos en decisiones automatizadas.
- Utilización de datos complementarios o generación de datos sintéticos que representen a estos colectivos.

Con este tipo de actuaciones, se cumple con la normativa vigente al mismo tiempo que se previene la exclusión desde el origen del sistema de IA; se incorporan voces diversas en la definición de variables relevantes para mejorar la representatividad de estos colectivos y se mejora la empleabilidad de las personas en situación de vulnerabilidad o exclusión.

1.4.- Consideraciones éticas sobre los sesgos en la IA

El uso de la inteligencia artificial en Recursos Humanos presenta riesgos éticos significativos, especialmente en relación con los sesgos que pueden surgir en los modelos algorítmicos. Estos sesgos pueden ser el resultado de datos históricos, decisiones de diseño y la interacción humana con los sistemas de IA.

Los sesgos en las herramientas de IA que pueden proceder de desigualdades históricas, estereotipos y errores en el diseño de modelos, pueden generar un efecto discriminatorio sobre grupos específicos. Partiendo de la premisa de que la imparcialidad absoluta es difícil de lograr debido a la subjetividad “humana” en la valoración de criterios de éxito, es razonable pensar que los modelos de IA reproducirán las desigualdades estructurales de la sociedad.
Diversas investigaciones determinan que los sesgos en la IA provienen de tres dimensiones principales: datos, modelo y uso. Cada una de estas dimensiones tiene implicaciones directas en la toma de decisiones en Recursos Humanos. Algunos ejemplos:

- Los datos sesgados pueden incluir prejuicios históricos que afectan las decisiones automatizadas.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

- El sesgo en el algoritmo puede surgir de decisiones técnicas en el diseño y en el entrenamiento de los modelos.
- La implementación y uso de la IA pueden amplificar sesgos existentes a través de bucles de aprendizaje.

Con el objetivo de garantizar un uso ético de la IA en el ámbito de los Recursos Humanos, es crucial garantizar que los empleados y candidatos comprendan cómo se toman las decisiones que les afectan, que el modelo sea explicable y permita identificar y corregir posibles sesgos y establecer protocolos de rendición de cuentas que asigne responsabilidades en caso de errores en las decisiones tomadas con ayuda de un algoritmo automatizado.

La sensibilización sobre el uso de IA es clave para promover una cultura organizativa crítica y participativa. Una comunicación clara sobre los sistemas algorítmicos, complementada con un plan de formación efectivo, refuerza la inclusión digital y la democratización tecnológica.

Es necesario desplegar planes de comunicación sobre la IA que sean accesibles y adaptables a los diferentes niveles de la organización. Más allá de informar, la comunicación debe conseguir la participación activa y la inquietud por saber. Los portales de explicabilidad, por ejemplo, son herramientas clave para facilitar el acceso a información sobre los algoritmos implementados en una organización.

Una vez que se genera la inquietud, las personas trabajadoras deben tener acceso a formación específica sobre el uso ético y responsable de la IA en el ámbito laboral. Una de las claves es diseñar un ecosistema formativo que involucre a toda la organización, de carácter transversal y adaptada a los diferentes roles; con contenidos clave como la identificación de sesgos algorítmicos y los derechos digitales y que incluya metodologías interactivas y espacios deliberativos. Son pocas las empresas, tanto en el ámbito privado como el público, que trabajan con este enfoque.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

2.- MARCO LEGAL PARA IMPLANTAR IA EN RR. HH

La integración de la inteligencia artificial (IA) en recursos humanos (RRHH) presenta oportunidades y desafíos significativos en términos de legalidad, ética y protección de datos. Es crucial establecer un marco normativo que garantice la equidad y la inclusión en el uso de estas tecnologías.

El cuerpo legal actual en Europa y en España es analizado en detalle en el Libro Blanco, poniendo el foco en los retos de cumplimiento que plantea la IA en el trabajo. El texto ofrece una lectura integrada del AI Act, el RGPD, la DSA, el Data Act y otros marcos regulatorios relevantes, con especial atención a sus implicaciones prácticas en la gestión de datos laborales, la toma de decisiones a partir de los resultados de los algoritmos y los derechos fundamentales de las personas trabajadoras.

En este resumen ejecutivo recogemos las cuestiones que consideramos más relevantes de la normativa europea y española, por lo que emplazamos a los interesados en profundizar en este ámbito o ampliar la perspectiva fuera del marco de la Unión Europea, a consultar el Libro Blanco y las actualizaciones que se vayan haciendo en base a los desarrollos normativos.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

2.1. El RGPD y el consentimiento informado

El RGPD establece obligaciones sobre el consentimiento informado y la transparencia en el uso de datos personales. El consentimiento informado es esencial para el uso de IA en los procesos de gestión del talento por lo que las empresas deben garantizar la transparencia y esto incluye informar a los candidatos sobre el uso de algoritmos, los datos utilizados y los criterios de selección de dichos datos.

Además, la normativa establece que tanto los empleados como los candidatos, deben otorgar su consentimiento explícito para el uso de sistemas de IA. La falta de este consentimiento puede resultar en multas de hasta 20 millones de euros o el 4% de la facturación anual.

La prevención de sesgos es crucial para garantizar la equidad en el uso de IA en RRHH. Los algoritmos deben ser diseñados y auditados para evitar la discriminación basada en características protegidas; y se deben llevar a cabo evaluaciones de impacto en protección de datos (EIPD) para identificar posibles riesgos. El RGPD prohíbe expresamente las decisiones automatizadas sin supervisión humana que afecten, significativamente, a los individuos.

La LOPDGDD complementa el RGPD e introduce derechos digitales específicos que protegen la privacidad y los derechos de los trabajadores en el uso de IA. Estos derechos garantizan un entorno laboral justo y equitativo.

2.2.- El Estatuto de los Trabajadores y el equilibrio entre la vida laboral y personal

El Artículo 20 bis del Estatuto de los Trabajadores establece derechos fundamentales para los trabajadores en relación con el uso de dispositivos digitales y la desconexión digital. Este marco legal busca proteger la intimidad y garantizar un equilibrio entre la vida laboral y personal. De esta forma, establece que las empresas deben garantizar que las políticas sobre el uso de dispositivos y desconexión sean claras y consensuadas. Además, deben evitar

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

prácticas que vulneren la intimidad de los trabajadores, como el monitoreo excesivo de correos electrónicos y la actividad en dispositivos corporativos.

Así mismo establece que:

- Los trabajadores tienen derecho a la intimidad en el uso de dispositivos digitales proporcionados por la empresa.
- El acceso a los contenidos de estos dispositivos por parte del empleador debe ser proporcional y necesario.
- Se prohíbe el uso de sistemas de vigilancia para fines distintos a los establecidos y siempre respetando la normativa de protección de datos.
- Los trabajadores tienen derecho a no estar conectados fuera de su horario laboral, salvo excepciones justificadas.
- Las empresas deben elaborar políticas internas de desconexión digital en consulta con los representantes de los trabajadores.

Además, el Artículo 20 bis exige que las decisiones automatizadas en el ámbito laboral cuenten con supervisión humana y que se informe a los trabajadores sobre los sistemas de IA que afectan sus condiciones laborales. Los trabajadores deben conocer los parámetros y reglas de los sistemas de IA utilizados en la toma de decisiones laborales; las decisiones críticas, como despidos o promociones, deben ser revisadas por un supervisor humano y las empresas deben informar sobre la lógica aplicada en decisiones automatizadas. Las legislaciones española y europea establecen un marco normativo para la implementación de IA en Recursos Humanos, priorizando la protección de derechos fundamentales y la equidad.

- La Ley 12/2021 (denominada Ley Ryder) adapta la legislación laboral al contexto de plataformas digitales, estableciendo una presunción de laboralidad para los repartidores que actuaban como autónomos.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

- La Agencia Española de Supervisión de la Inteligencia Artificial (AESIA) supervisa el desarrollo y uso de la IA en España, garantizando el cumplimiento de normativas como el RGPD y la LOPDGDD.

- La Carta de Derechos Digitales promueve principios de igualdad, no discriminación y transparencia en el uso de IA.

2.3 Estrategias de Cumplimiento Normativo

El cumplimiento normativo debe ser un componente estructural en el diseño organizativo de la IA. Para mitigar los posibles problemas legales es necesario identificar, de forma temprana, los riesgos que supone su uso, como la discriminación indirecta, la vulneración de la intimidad o la falta de supervisión humana. Estos riesgos pueden ir asociados a la desconfianza de los empleados y el incremento de la litigación laboral, si se percibe falta de transparencia.

Las estrategias eficaces no solo evitan sanciones, sino que también fomentan una cultura ética. Estas son algunas recomendaciones:

- Establecer protocolos de validación y revisión continua para asegurar la equidad en los resultados. Esto incluye la realización de auditorías y la designación de un delegado de Protección de Datos.

- Realizar EIPD antes de implementar sistemas de IA que impliquen tratamientos de datos sensibles.

- Desarrollar políticas internas de uso responsable de IA que permita definir límites y responsabilidades.

- Los algoritmos deben ser diseñados para evitar sesgos y garantizar igualdad de oportunidades en procesos de selección. La normativa prohíbe expresamente la discriminación directa o indirecta en el empleo por motivos de edad, discapacidad, sexo, origen, religión, entre otros.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Las buenas prácticas en la gestión de riesgos derivados de la IA son fundamentales para convertir principios éticos en procesos sostenibles. Estas prácticas permiten anticipar problemas y mejorar la calidad de las decisiones.

3.- INVESTIGACIÓN EMPÍRICA DE IA EN RR. HH

El proyecto IA+Igual plantea el análisis de herramientas de IA aplicadas a procesos del ámbito laboral con el objeto de determinar cómo los sesgos pueden estar generado riesgos de discriminación.

El proyecto propone un marco de análisis multidisciplinar robusto, cuantitativo y cualitativo, que integra cuatro dimensiones: técnica, legal, ética y RR. HH., en la implementación de estas tecnologías.

- La Gobernanza del dato es fundamental para la trazabilidad y control de la IA en RR. HH.
- Normativas como el AI Act y el GDPR establecen requisitos de calidad y representatividad de los datos.
- Roles como Data Owners y Data Stewards son cruciales para la gestión de datos.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

- Se requiere un control técnico exhaustivo a lo largo del ciclo de vida del dato.

- Herramientas como Warden AI permiten auditorías continuas de sesgos en procesos de RRHH.

El valor añadido y diferencial de este proyecto es haber realizado este análisis con un enfoque empírico que ha permitido detectar el impacto de los sesgos en un entorno real. El equipo IA+Igual ha analizado 7 algoritmos aplicados a procesos de RR. HH., en diferentes empresas y 3 buenas prácticas relacionadas con la Gobernanza del dato.

Una mirada comparativa: riesgos y medidas desde cada dimensión

Dimensión	**Riesgo detectado**	**Medida correctiva propuesta**
Técnica	Modelo sobreajustado o inestable	Validación cruzada, revisión continua del rendimiento
Legal	Tratamiento desproporcionado de datos personales	Test de proporcionalidad, intervención humana significativa
Ética	Reproducción de sesgos históricos	Evaluación de equidad algorítmica, explicabilidad
RRHH	Resultados no alineados con la cultura y estrategia	Supervisión funcional, contextualización organizativa

Hablar de sesgo algorítmico no significa que los modelos “opinen” o “discriminen” como lo haría una persona. Significa que, como herramientas estadísticas entrenadas sobre datos del mundo real, los algoritmos pueden reproducir, amplificar o encubrir desigualdades existentes, sin ser conscientes de ello ni poder corregirse por sí mismos.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Un sesgo algorítmico puede definirse como una distorsión sistemática e injustificada en el comportamiento de un sistema automatizado que afecta de forma diferencial a ciertos grupos o individuos. No es lo mismo que un error aleatorio: se trata de un patrón persistente, predecible y estructural (Mehrabi et al., 2021).

Este tipo de sesgos puede tener consecuencias graves, especialmente en contextos sensibles como el ámbito laboral. La IA en RRHH puede tomar decisiones que influyen directamente en las trayectorias profesionales de las personas: acceso al empleo, promociones, despidos, evaluaciones o salarios. Si esas decisiones están sesgadas, los efectos no solo son injustos, sino que pueden perpetuar o institucionalizar discriminaciones históricas.

El sesgo no aparece únicamente “en los datos” ni “en el algoritmo”. Es un fenómeno multicausal que puede surgir en cualquiera de las siguientes fases:

- Sesgo en los datos de entrada: Por ejemplo, si el conjunto de entrenamiento está desbalanceado o refleja decisiones humanas previas que ya eran discriminatorias.
- Sesgo en la construcción del modelo: Ocurre cuando el diseño del sistema —elección de variables, funciones de coste, umbrales— favorece ciertos perfiles o penaliza a otros sin justificación técnica ni ética.
- Sesgo en la implementación y uso: Puede darse cuando el modelo se aplica en contextos para los que no fue validado, o cuando se combina con procesos organizativos que lo sesgan aún más.

Existen múltiples formas en las que el sesgo puede manifestarse. Algunas de las más relevantes en entornos de RRHH son:

- Sesgo de representación: Algunos grupos están infrarrepresentados en los datos, lo que reduce la precisión del modelo para esos perfiles.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

- Sesgo de selección: El modelo aprende patrones previos de selección que pueden haber sido discriminatorios.

- Sesgo de exclusión: Se omiten variables clave o se usan proxies que ocultan dimensiones sociales relevantes.

- Sesgo de medición: La variable objetivo (por ejemplo, "rendimiento" o "éxito") está mal definida o refleja valores cuestionables, como mostró el caso documentado por Obermeyer et al. (2019) en el ámbito sanitario.

- Sesgo de evaluación: Se aplican métricas globales sin desagregarlas por subgrupos, ocultando desigualdades estructurales.

Estas tipologías no son excluyentes. De hecho, suelen combinarse y retroalimentarse, generando distorsiones difíciles de detectar si no se analizan explícitamente.

Uno de los mayores peligros de los sistemas algorítmicos es que pueden aprender sesgos del pasado y proyectarlos hacia el futuro. Si no se interviene, esto genera un ciclo vicioso:

1. **Datos históricos** con desigualdades.
2. **Modelo entrenado** con esos datos.
3. **Resultados sesgados** aplicados a nuevas personas.
4. **Retroalimentación** que refuerza los patrones previos.

3.1.- Análisis cualitativo

Los análisis cualitativos son fundamentales para comprender el impacto de la IA en Recursos Humanos. Capturar percepciones y experiencias de los usuarios permite complementar los datos cuantitativos y mejorar la legitimidad social de la IA, porque ayudan a identificar inquietudes y resistencias ante su implantación. Herramientas como NVivo y ATLAS.ti facilitan la codificación y visualización de patrones en este ámbito

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Desde este punto de vista cualitativo, la evaluación de la equidad en sistemas algorítmicos es crucial para identificar y mitigar sesgos.

3.2- Análisis cuantitativo

Una vez comprendido qué es el sesgo y cómo puede surgir, el siguiente paso es determinar cómo evaluarlo de forma empírica y cuantificable. Para ello, el campo de la inteligencia artificial ha desarrollado un conjunto de métricas llamadas fairness metrics, que buscan traducir distintas nociones de equidad en términos matemáticos. Estas métricas permiten diagnosticar cuándo un sistema automatizado está generando resultados injustos o desiguales, y en qué medida.

Sin embargo, medir la equidad no es una tarea neutral ni sencilla. No existe una única definición de justicia, ni todas las aproximaciones son compatibles entre sí. Según la filosofía política y la teoría de la probabilidad, las métricas de equidad pueden agruparse en tres familias conceptuales: independencia, separación y suficiencia.

- **Independencia (Paridad Demográfica):** Parte de la idea de que la decisión del algoritmo debe ser independiente de atributos sensibles como género, edad o etnia. Su métrica más conocida es la paridad demográfica, que exige que todos los grupos reciban decisiones positivas en la misma proporción. Este enfoque está vinculado a ideas de igualdad de trato formal y de no discriminación directa. Aunque fácil de calcular, suele ser cuestionado cuando existen diferencias reales en los perfiles de desempeño entre grupos. Se busca que las decisiones del algoritmo sean independientes de atributos sensibles como género, edad o etnia.

- **Separación (Oportunidad Igualitaria):** La separación propone que el algoritmo puede tomar decisiones distintas para distintos grupos, pero no debe cometer más errores con unos que con otros. Por eso estas métricas requieren que la predicción esté condicionada al resultado real. Lo

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

importante es que el modelo cometa errores de forma similar entre grupos. Está alineada con la idea de igualdad de acceso a oportunidades. Su implementación suele preferirse en contextos donde se busca minimizar desigualdades de trato en quienes sí cumplen los criterios objetivos de éxito.

- **Suficiencia (Paridad Predictiva):** exige que la decisión del modelo sea igualmente confiable para todos los grupos. Es decir, si un modelo predice que una persona será apta, esa predicción debería tener igual validez sin importar su grupo.

Sin embargo, no es posible cumplir simultáneamente con los tres criterios de equidad cuando existen diferencias reales entre grupos. La elección de una métrica implica decisiones normativas y debe ser transparente.

- **Tríada Imposible:** Se ha demostrado que no se pueden cumplir los tres criterios de equidad al mismo tiempo.

- **Ejemplos de Elección:** Un sistema de justicia puede optar por separación, mientras que una empresa puede preferir suficiencia.

- **Transparencia:** Es fundamental documentar el criterio adoptado para la evaluación de la equidad.

Este marco conceptual es clave para interpretar correctamente las métricas cuantitativas de fairness, y será la base sobre la que se construya la comparación entre tipos de sesgo, métricas concretas y mitigaciones.

Comprender los distintos tipos de sesgo y los enfoques formales de equidad es indispensable, pero no suficiente para intervenir con eficacia. En la práctica, los equipos técnicos y responsables de análisis algorítmico necesitan traducir esas nociones en acciones concretas. Incluimos una tabla que relaciona tipos de sesgo, métricas de detección y técnicas de mitigación. Esta herramienta ayuda a orientar los análisis algorítmicos y a tomar decisiones técnicas.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Técnicas de mitigación y tipos de sesgos

Tipo de sesgo	Métrica de detección	Técnica o enfoque de mitigación
Sesgo de representación	Paridad demográfica (*demographic parity*), Skew ratio (Verma & Rubin, 2018)	Reponderación de clases, oversampling de grupos infrarrepresentados (Kamiran & Calders, 2012)
Sesgo de selección	*Equal opportunity*, *Equalized odds* (Hardt, Price & Srebro, 2016)	Ajuste del clasificador, regularización de errores condicionales (Friedler et al., 2019)
Sesgo de exclusión	Fairness Through Awareness (Dwork et al., 2012), análisis de proxies (Barocas et al., 2019)	Inclusión de variables latentes, análisis causal, revisión del feature set
Sesgo de medición	*Predictive parity*, análisis de fiabilidad (Chouldechova, 2017)	Redefinición de la variable objetivo, supervisión humana, modelado multitarea
Sesgo de evaluación	Disparidad en métricas por subgrupo (accuracy, recall, F1) (Verma & Rubin, 2018)	Evaluación estratificada, dashboards desagregados, revisión de umbrales
Sesgo interseccional	Disparate Impact Ratio por subgrupos compuestos (Kearns et al., 2019)	Mitigadores multietapa, optimización de Pareto, fairness individual
Sesgo de confirmación (epistémico)	Revisión del pipeline, análisis de dependencia excesiva (Holstein et al., 2019)	Documentación de hipótesis, validación externa, participación interdisciplinar

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Las estrategias de mitigación se clasifican en tres momentos del ciclo de vida del modelo: preprocesamiento, in-training y postprocesamiento.

- **Preprocesamiento:** Modifica datos antes del entrenamiento para neutralizar sesgos, como balanceo estadístico.
- **In-training:** Incorpora restricciones durante el entrenamiento para penalizar desigualdades, como fairness constraints.
- **Postprocesamiento:** Ajusta decisiones del modelo después del entrenamiento, como calibración de resultados.

Se proponen estrategias multietapa para abordar el sesgo algorítmico en diferentes fases del ciclo de vida del modelo.

- **Mitigación Previa:** Actúa desde el diseño de datos y variables.
- **Mitigación Local:** Focalizada en puntos críticos como umbrales y decisiones individuales.
- **Análisis Interseccional:** Desagregar métricas por subgrupos compuestos para detectar desigualdades ocultas.

Además, es esencial contar con herramientas que permitan evaluar y visualizar la equidad en los modelos de IA.

- **SHAP y LIME:** Herramientas para interpretar decisiones individuales de modelos complejos
- **AI Fairness 360 y Fairlearn:** Bibliotecas que ofrecen métricas de fairness y técnicas de mitigación.
- **Dashboards de Equidad:** Permiten observar dinámicamente cómo varían las métricas por grupo.

Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO DE DERECHOS SOCIALES, CONSUMO Y AGENDA 2030
Plan de Recuperación, Transformación y Resiliencia
Comunidad de Madrid
Dirección General de Evaluación, Calidad e Innovación
CONSEJERÍA DE FAMILIA, JUVENTUD Y ASUNTOS SOCIALES

4.- ESTÁNDAR DE CERTIFICACIÓN DE IA LABORAL

Se propone un sistema de certificación para la IA en recursos humanos que reconozca el compromiso ético y técnico de las organizaciones. Este sistema se basa en un enfoque escalonado que permite a las entidades avanzar en su madurez ética. Dicho sistema parte del desarrollo de un sello ético, una herramienta de transformación cultural, que fomente la alfabetización algorítmica y la creación de comités éticos en las organizaciones.

El sello ético se fundamenta en los principios de justicia, transparencia y responsabilidad, alineados con marcos internacionales de ética digital. Además, su desarrollo debe estar alineado con normas internacionales consolidadas para asegurar su credibilidad. Esto facilitará la interoperabilidad con auditorías externas y requisitos legales.

4.1.- Enfoque Progresivo para la Mejora Continua

Se propone un modelo de certificación escalonado que permita a las organizaciones avanzar según su alineamiento con principios éticos. Este enfoque refuerza la idea de que la certificación es un proceso evolutivo.

- Niveles de certificación: Bronce, Plata, Oro y Platino.
- Cada nivel tiene requisitos específicos que reflejan el compromiso ético.
- Se busca fomentar una lógica de mejora continua en el uso de IA.

Un sistema de certificación robusto debe considerar el ciclo de vida de los sistemas algorítmicos, evaluando datos, modelos y el contexto organizativo. Esto asegura un uso justo y responsable de la IA. Además, la inclusión de terceros en el proceso de certificación refuerza la credibilidad del sello. Esto incluye auditorías independientes y la participación de agentes sociales.

- Auditorías por entidades independientes certificadas.
- Creación de comités éticos internos o interdisciplinares.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

- Publicación de informes de transparencia sobre el impacto de la IA.

4.2. Riesgos en la Certificación Ética de IA

La implementación de un sistema de certificación ética en inteligencia artificial aplicada a recursos humanos enfrenta varios riesgos que pueden comprometer su efectividad y legitimidad. Es crucial abordar estos riesgos para asegurar un marco ético sólido y funcional.

- **Evitar la banalización de la ética:** El "certification washing" puede deslegitimar los esfuerzos reales, convirtiendo el sello ético en un símbolo vacío.

- **Requisitos exigentes:** Se deben establecer criterios proporcionados y verificables para evitar la superficialidad en la adopción de la ética.

- **Mecanismos de revocación**: Incluir la posibilidad de revocar el sello ante incumplimientos es esencial para mantener la credibilidad.

Además, es necesaria la participación de las administraciones públicas, cuyo papel es crucial en la promoción y legitimación del modelo de certificación ética en IA, actuando como facilitadoras y catalizadoras.

- **Incorporación del sello en licitaciones:** Utilizar el sello como referencia en compras públicas para garantizar el uso ético de la IA.

- **Políticas de incentivos:** Promover beneficios para organizaciones que obtengan niveles avanzados de certificación.

- **Programas de acompañamiento:** Ofrecer asesoramiento y formación técnica para facilitar la implementación del sistema.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

5.- CONCLUSIONES

El propósito de estas conclusiones es servir como base para la acción y como punto de partida para futuras políticas, estándares técnicos y líneas de investigación en torno a una IA laboral más ética, inclusiva y eficaz.

1. Sesgos en el diseño y uso de algoritmos en función de los Ejes de Actuación

Uno de los hallazgos más preocupantes del análisis es la presencia de sesgos no detectados en los sistemas analizados empíricamente. Muchas organizaciones, con la intención de evitar discriminaciones, eliminan variables como género o discapacidad. Sin embargo, esta práctica no garantiza la neutralidad, sino que puede ocultar sesgos implícitos que se mantienen a través de otras variables correlacionadas.

Además, ninguna de las empresas evaluadas realizó un análisis técnico riguroso para identificar sesgos potenciales. Se evidencia una falta de comprensión sobre cómo los sesgos pueden surgir de forma estructural y cómo deberían abordarse de forma proactiva, más allá de la simple omisión de variables sensibles.

El impacto de estos sesgos es especialmente grave en colectivos vulnerables, cuya invisibilización puede comprometer directamente sus oportunidades de empleabilidad.

2. Cumplimiento normativo y gobernanza del dato

Los análisis empíricos revelan un cumplimiento muy deficiente de la normativa vigente. Ninguna de las organizaciones realizó una evaluación de impacto previa, a pesar de tratarse de algoritmos de alto riesgo en el marco de la regulación europea (RIA). Además, se han identificado usos no autorizados de algoritmos más allá de su finalidad original, lo que constituye una infracción grave del principio de "limitación de la finalidad".

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

Este vacío normativo se agrava por una falta de gobernanza interna del dato: las empresas no disponen de estructuras ni procedimientos claros para gestionar el ciclo de vida de los datos ni para supervisar las implicaciones legales de sus sistemas automatizados.

3. Formación y cultura organizativa

Existe un fuerte desequilibrio entre la formación técnica (uso de herramientas como ChatGPT, Copilot, etc.) y la formación ética, crítica o estratégica. La mayoría de los empleados y directivos desconocen los riesgos reales asociados al uso de IA y carecen de recursos para evaluar adecuadamente su implementación.

Además, se observa una externalización excesiva del desarrollo de algoritmos, sin supervisión suficiente por parte de los responsables de RRHH. Esto contribuye a una cultura de dependencia tecnológica sin control, donde las decisiones quedan en manos de los técnicos sin una visión de conjunto.

4. Inclusión social y colectivos invisibilizados

La eliminación de variables sensibles ha llevado a la invisibilización de colectivos vulnerables, como personas con discapacidad, migrantes o víctimas de violencia de género. Al no registrar datos sobre condiciones relevantes, se impide realizar análisis que identifiquen desigualdades estructurales.

Se proponen alternativas como el uso de datos sintéticos para representar estos perfiles de forma segura. También destacan enfoques innovadores como describir "capacidades funcionales" en lugar de "discapacidades", lo que abre nuevas posibilidades para la inclusión desde el diseño algorítmico.

Estas estrategias contribuyen directamente a la mejora de la empleabilidad, autonomía y protección de los derechos de estos colectivos.

Financiado por la Unión Europea NextGenerationEU
Gobierno de España
Ministerio de Derechos Sociales, Consumo y Agenda 2030
Plan de Recuperación, Transformación y Resiliencia
Comunidad de Madrid
Dirección General de Evaluación, Calidad e Innovación
Consejería de Familia, Juventud y Asuntos Sociales

5. Recursos Humanos: debilidades en la compra y uso de IA

Se ha constatado una falta generalizada de preparación para evaluar, seleccionar e implementar sistemas de IA. No se realizan análisis de impacto, no se consideran alternativas menos invasivas, y no se exige a los proveedores el cumplimiento de criterios de explicabilidad o proporcionalidad. El uso de modelos excesivamente complejos, sin justificación técnica clara, añade opacidad al proceso. En muchos casos, soluciones más simples y explicables habrían bastado. Esta tendencia refleja un desajuste entre el entusiasmo por la IA y la comprensión real de sus implicaciones prácticas.

6. Certificación ética e implementación progresiva

Una conclusión clave es la necesidad de establecer mecanismos de certificación ética adaptados al grado de madurez de cada organización. Se propone un modelo escalable e incremental, que empiece con requisitos básicos como la existencia de formación específica, comités éticos internos y procesos mínimos de documentación. Esta certificación debería fomentar una cultura ética desde la base, ofreciendo seguridad jurídica y metodológica a las organizaciones, y facilitando su avance hacia modelos de IA más seguros, explicables y socialmente responsables.

7. Reflexiones finales: desconocimiento y riesgo

Muchas organizaciones aplican IA sin saber realmente cómo funciona ni qué consecuencias puede tener. Se utilizan modelos complejos innecesarios, difíciles de explicar, cuando en muchos casos bastarían métodos más simples y comprensibles.

Esta incertidumbre genera dos tipos de reacción: paralización (frenar la innovación por miedo) o inconsciencia (aplicar IA sin evaluación). Ambas respuestas son inadecuadas y comprometen la equidad, la legalidad y los derechos fundamentales.

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

En el ámbito de la innovación social, esto representa un triple reto:

- Cualificar a los profesionales para el uso ético de IA
- Reconocer a los colectivos vulnerables en los datos
- Impedir que la IA agrave la desigualdad o perjudique los ODS

Tabla resumen de conclusiones

Eje temático	Problema observado	Conclusión principal	Recomendación
Sesgos	Eliminación de variables sensibles sin análisis de fondo	Los sesgos implícitos persisten	Realizar auditorías de sesgos con enfoque estructural
Normativa y gobernanza	No se hacen evaluaciones de impacto ni se controla el uso de algoritmos	Incumplimiento sistemático de la RIA y falta de gobernanza	Aplicar evaluaciones de impacto y establecer control interno de algoritmos
Formación	Formación técnica sin perspectiva crítica	Ausencia de cultura ética y estratégica	Integrar pensamiento crítico y formación ética en todos los niveles
Inclusión	Invisibilización de colectivos vulnerables	Se impide detectar desigualdades reales	Usar datos sintéticos y enfoques centrados en capacidades
Recursos Humanos	Desconocimiento en la compra y aplicación de IA	RRHH no sabe evaluar ni supervisar algoritmos	Generar guías internas y formación para responsables de RRHH
Certificación	No hay criterios unificados ni escalabilidad	Falta de confianza y miedo generalizado	Promover certificación escalable desde la sensibilización hasta la auditoría
Reflexiones finales	Desconocimiento del funcionamiento y riesgos de la IA	Uso inadecuado de IA o inmovilismo por miedo	Fomentar cultura de innovación responsable y evaluación multidisciplinar

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

CONTACTO

Marisa Cruzado
626171799
www.iamasigual.eu
cva@cvalora.com

Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia