Una nueva agenda de gobernanza algorítmica para la negociación colectiva

A New Agenda for Algorithmic Governance in Collective Bargaining

María Villa Fombuena
Profesora Ayudante Doctora Derecho del Trabajo y de la Seguridad Social
(Acreditada Profesora Titular)
Universidad de Sevilla, España
mvilla1@us.es
ORCID: 0000-0002-0192-5051

Resumen: La regulación vigente sobre algoritmos laborales descansa sobre la ilusión de que la transparencia es suficiente. Sin embargo, informar a las personas trabajadoras sobre cómo funciona un sistema no les protege si carecen de poder real para influir en su diseño, validar su comportamiento, o impugnarlo cuando causa discriminación. La negociación colectiva debe asumir un rol radicalmente distinto: no limitarse a reaccionar ante sistemas ya implantados, sino ejercer una gobernanza activa del ciclo completo de vida de los algoritmos laborales. Se propone en consecuencia un modelo multifásico donde la representación de las personas trabajadoras interviene desde el diseño hasta el desmantelamiento, tratando de evitar la materialización de una paradoja peligrosa: el cumplimiento formal, sin equidad real. Brecha que sólo una evaluación de impacto laboral -específica, contextualizada, y negociable colectivamente- puede cerrar. El resultado es un modelo donde la tecnología está subordinada a los principios del Derecho del Trabajo, y no al revés.

Palabras claves: Gobernanza algorítmica; Negociación colectiva; Ciclo de vida del algoritmo; Derechos laborales fundamentales.

Abstract: Current regulations on workplace algorithms are based on the illusion that transparency is sufficient. However, informing workers about how a

Recibido: 9 diciembre 2025; Aceptado: 14 enero 2026

e-ISSN: 2660-4884 *Trabajo, Persona, Derecho, Mercado* 11 (2025) 25-49
https://dx.doi.org/10.12795/TPDM.2025.i11.01

system works does not protect them if they lack real power to influence its design, validate its behaviour, or challenge it when it causes discrimination. Collective bargaining must take on a radically different role: not merely reacting to systems that are already in place, but actively governing the entire life cycle of labour algorithms. We therefore propose a multi-phase model in which workers' representatives are involved from design to dismantling, seeking to avoid the emergence of a dangerous paradox: formal compliance without real equity. This gap can only be closed by a labour impact assessment that is specific, contextualised and collectively negotiable. The result is a model where technology is subordinate to the principles of labour law, and not the other way around.

Keywords: Algorithmic governance; Collective bargaining; Algorithm lifecycle; Fundamental labour rights.

1. La gobernanza algorítmica laboral. Necesidad de una concepción multifásica

La irrupción de los algoritmos en la gestión del trabajo ha desbordado el marco tradicional de la transparencia, exigiendo a la negociación colectiva y al Derecho del Trabajo una nueva arquitectura de control sobre herramientas que configuran condiciones, trayectorias y bienestar profesional a través de todo el ciclo vital laboral. En este contexto, la gobernanza algorítmica laboral debe concebirse no meramente como la obligación de informar sobre la existencia de sistemas automáticos, sino como la constitución de un subsistema jurídico, en el que reglas, procedimientos y órganos paritarios regulan el diseño, entrenamiento, implantación, seguimiento, revisión y eventual retirada de algoritmos que afectan a decisiones laborales esenciales. En otras palabras, no basta con informar; hay que ir más allá de la mera transparencia para articular reglas, procedimientos y órganos de control sobre el ciclo completo de vida del algoritmo aplicado a la gestión laboral.

Desde esta óptica, la gobernanza algorítmica del trabajo debería entenderse como el conjunto de reglas internas de la empresa, pactadas o condicionadas por la negociación colectiva, que constituyen el algoritmo a aplicar (qué decisiones puede adoptar, con qué límites, qué datos puede usar y con qué correcciones *ex ante*). Esta gobernanza convierte al algoritmo en un subsistema normativo sometido a los principios del Derecho del Trabajo (protección de la parte débil, igualdad, salud laboral, participación colectiva), y no al revés.

Es por ello por lo que la negociación colectiva y la representación de las personas trabajadoras (en adelante, RLT) se han de situar en el centro de los procesos críticos del ciclo de vida del algoritmo aplicado al trabajo.

Haciendo una adaptación propia de la lógica "iniciousocierre" que ponen de manifiesto los principales estándares o marcos generales sobre ciclos de vida de Inteligencia Artificial (en adelante, IA) y datos[1], y llevando a cabo una reordenación (igualmente propia), que permite que cada fase marque un momento diferenciado de intervención de la negociación colectiva y de la RLT, podrían diferenciarse cinco fases o etapas básicas en el ciclo vital de toda IA aplicada al ámbito laboral:

Una primera fase de diseño en la que se va a constituir el algoritmo laboral: qué puede hacer, con qué datos y con qué límites. Etapa que aglutina por tanto tres elementos estratégicos:

— Finalidad del sistema: para qué se va a utilizar (selección de personal, cribado curricular, asignación de turnos y rutas, evaluación del desempeño, cálculo de incentivos, control horario, detección de incumplimientos, etc).
— Datos y variables: fuentes de datos con las que se cuenta (bases internas de recursos humanos, historiales disciplinarios, datos biométricos, información extraída de redes sociales, geolocalización, *wearables*, etc.); condiciones en las que se recaban; de esos datos, qué atributos se van a usar y cómo se ponderan y qué riesgos de sesgo o invasión de la privacidad comportan.
— Límites materiales y garantistas al algoritmo: qué no puede hacer, qué decisiones no puede automatizar en exclusiva, qué datos no puede utilizar y con qué salvaguardas mínimas de intervención humana debe contar.

1. La referencia técnica central en materia de "ciclo de vida" de sistemas de IA con aprendizaje automático es el estándar ISO/IEC 23053- Framework for AI systems using machine learning, que se cita de forma recurrente en documentos de gobernanza de IA y en materiales de estandarización europeos como referencia básica para ordenar procesos y responsabilidades a lo largo de todo el ciclo del sistema. Sirva de ejemplo el documento "CENCENELEC JTC 21 AI Standards: Complete Detailed Overview", cuya función principal es servir de referencia para reguladores, industria y otros actores a la hora de saber qué estándares técnicos pueden utilizarse o desarrollarse como apoyo a la implementación del Reglamento Europeo de Inteligencia Artificial y a la construcción de un ecosistema europeo de IA segura y fiable, y que incluye la ISO/IEC 23053 dentro del repertorio de normas europeas relevantes para la IA.

En una segunda etapa, que podría denominarse fase de entrenamiento y pruebas, la gobernanza algorítmica laboral se desplaza desde el plano de las decisiones de diseño (plano teórico o de la intención) al de la verificación empírica de sus consecuencias (plano fáctico o de los hechos). El sistema ya no es sólo un proyecto, sino que ha sido alimentado con datos y empieza a generar predicciones y decisiones, eso sí, simuladas como se verá más adelante. Es en este momento cuando se adquiere verdadera conciencia del modelo que después operará sobre decisiones en el ámbito laboral. Desde la perspectiva jurídica, se trata de una etapa trascendental porque es precisamente aquí donde van a emerger los posibles sesgos y errores que derivan de los datos con los que se ha alimentado el sistema; como consecuencia de lo cual, la gobernanza algorítmica laboral exige que, como en la etapa anterior, esta fase se abra también a la intervención informada de la representación de las personas trabajadoras.

Sin embargo, antes de que el sistema se generalice más allá del piloto, la gobernanza planteada aquí exige que se realice una evaluación de impacto específicamente diseñada para el contexto laboral concreto y la protección de derechos del colectivo afectado. Esta evaluación, que es distinta de la evaluación de impacto sobre los derechos fundamentales exigida por el artículo 27 del Reglamento europeo sobre IA, aunque complementaria, como se verá más adelante, constituye una tercera fase o etapa trascendental en la que las evidencias recopiladas en la fase anterior van a quedar sistematizadas.

El siguiente hito se materializa en la fase de entrenamiento y pruebas, que constituye el primer momento de confrontación real entre el algoritmo y los principios del Derecho del Trabajo. No es una etapa puramente técnica reservada a la ingeniería de datos, sino un espacio donde la evidencia empírica sobre el comportamiento del modelo se examina bajo el escrutinio de quien tiene responsabilidad de proteger derechos laborales. En esta etapa, la participación de la RLT, articulada a través de acceso a información sobre comportamiento del modelo, participación en pilotos controlados, análisis de impacto compartido y capacidad de proponer medidas de mitigación, permite detectar y corregir sesgos antes de que el sistema se consolide en la empresa.

El ciclo de vida del sistema quedaría adecuadamente cerrado con una última fase: su desmantelamiento o sustitución. Etapa frecuentemente olvidada, pero que, como se va tener ocasión de examinar, es crítica para la seguridad jurídica de las personas trabajadoras.

2. Intervención negocial desde el diseño

Desde la perspectiva laboral, esta primera fase es crítica porque pueden programarse ya posibles discriminaciones (género, edad, origen, discapacidad) y desequilibrios de

poder que luego resultarán muy difíciles de corregir cuando el sistema esté en producción. Aquí es donde la gobernanza algorítmica laboral debe ser más intensa por tanto y debe producirse la entrada de la labor representativa (antes de que exista un modelo operativo). Entrada que debe materializarse a través de tres mecanismos conectados: el acceso a información temprana y cualificada, la capacidad de formular objeciones sustantivas sobre el diseño, y la incorporación de salvaguardas jurídicas mediante la negociación colectiva.

En primer lugar, el derecho de información debe ir significativamente más allá de lo que establece el artículo 64.4.d) del Estatuto de los Trabajadores, que reconoce la obligación empresarial de comunicar parámetros, reglas e instrucciones de los algoritmos ya implementados. La gobernanza algorítmica laboral que aquí se aborda exige que la RLT acceda a información estructurada y disponible en el momento mismo en que se están definiendo los contornos del sistema, es decir, antes de que las decisiones se cierren técnicamente. Esta información debe abarcar la finalidad concreta perseguida, los colectivos específicamente afectados, el tipo y alcance de las decisiones que se pretenden automatizar o semiautomatizar, los escenarios previstos de uso y los impactos esperados sobre dimensiones nucleares del empleo como la disponibilidad de oportunidades, la jornada, la remuneración y la construcción de carreras profesionales. Aspectos que no son marginales o meramente procedimentales, sino el sustrato mismo de lo que significa ser una persona trabajadora. Un algoritmo que afecte adversamente a todas o a algunas de ellas de manera simultánea, puede transformar radicalmente la experiencia y las oportunidades de vida de un/a trabajadoro/a. Por eso la representación debe conocer *ex ante* cuál será el impacto previsible del algoritmo en cada una de esas dimensiones. No se trata sólo de saber cómo funciona técnicamente, sino de comprender qué consecuencias reales tendrá en la vida laboral de las personas a las que afecta[2].

Esa información temprana debe complementarse con la capacidad efectiva de la representación laboral de formular objeciones jurídicas y organizativas de alcance general. En otras palabras, los y las representantes no sólo deben recibir información sobre el algoritmo, sino que deben tener capacidad real de objetar aspectos del diseño antes de que el sistema se implante. Esta capacidad de alerta temprana responde a dos problemas que la literatura y la evidencia empírica documentan de forma recurrente[3].

2. En línea con la Guía práctica del Ministerio de Trabajo y Economía Social sobre información algorítmica en el ámbito laboral (2022), que establece expresamente que la información que debe facilitarse a la representación legal de los trabajadores debe ser "significativa", lo cual implica que debe explicar la lógica del sistema, qué datos usa, qué variables son más relevantes y, qué impacto puede tener sobre las condiciones laborales. Es decir, la propia Guía oficial va más allá de exigir una descripción técnica del algoritmo: exige que se informe sobre las consecuencias reales en la vida laboral de las personas afectadas.

3. Véase, Fabregat Monfort, Gemma. "Procesos de selección algorítmica y discriminación", *LABOS*, Revista de Derecho del Trabajo y Protección Social, Vol.5, noviembre, 2024, pp. 131-153; y Ginès i

En primer término, resulta necesario identificar y cuestionar objetivos ilícitos o desproporcionados implícitos en el diseño del algoritmo, como puede ser la utilización encubierta del sistema como instrumento de sanción automatizada, la intensificación sin límites del control sobre la actividad laboral, o la invasión del derecho fundamental a la intimidad personal. En segundo término, y vinculado al anterior, es imprescindible escrutar qué información usa el algoritmo para tomar decisiones y si esa información puede generar discriminación, ya sea de manera directa (mediante atributos como el género, la edad, el origen racial o étnico, discapacidad, estado de salud o situación familiar) o a través de *proxies* de manera encubierta (este sería el caso, por ejemplo, si el algoritmo usa la dirección de la persona candidata para filtrar, y esa dirección correlaciona con barrios mayoritariamente habitados por personas de determinadas etnias. En este caso, se estaría discriminando indirectamente por origen racial)[4].

Esta capacidad de formular objeciones o de alerta temprana encuentra sustento tanto en la amplia evidencia científica de que los sistemas de IA para reclutamiento y gestión del personal discriminan si no se controlan desde el diseño[5], como en el deber empresarial preventivo de evaluar el impacto sobre derechos fundamentales antes del despliegue (art. 27) y de gestionar riesgos durante todo el ciclo de vida (art.

Fabrellas, Anna. "Analítica de personas y discriminación algorítmica en procesos de selección y contratación", LABOS, Revista de Derecho del Trabajo y Protección Social, Vol.5, noviembre, 2024, pp. 99-130.

Por lo que a evidencia empírica se refiere, el caso más conocido quizás sea el de Amazon, cuyo algoritmo de selección de personal para puestos técnicos penalizaba sistemáticamente currículums que incluían la palabra women's (por ejemplo, *women's chess club captain*) y rebajaba puntuaciones de graduadas de universidades femeninas; a lo que se unía además la circunstancia de que el algoritmo fue entrenado con currículums recibidos durante 10 años, mayoritariamente de hombres, reflejo de la composición del sector tecnológico, lo que tuvo como consecuencia directa la perpetuación de la infrarrepresentación histórica de mujeres en el sector tecnológico. Véase, Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women", *Reuters*, 10 de octubre de 2018. Disponible en: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/

4. Una variable proxy es una variable aparentemente neutral, que, en realidad, establece una correspondencia con una característica protegida por la ley (como puede ser el género, la edad, el origen étnico, la discapacidad, etc.). En otras palabras, es un dato que no pregunta directamente sobre una característica discriminatoria, pero que de facto funciona como un indicador indirecto de ella. Véase Prince, Anya and Schwarcz, Daniel. *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, Iowa Law Review, Vol. 105, 2020, pp.1257-1318.

Disponible en: https://ilr.law.uiowa.edu/sites/ilr.law.uiowa.edu/files/2023-02/Prince_Schwarcz.pdf

5. Además de los casos evidenciados por Fabregat Monfort, Gemma. "Procesos de selección algorítmica y discriminación", op. Cit.; Ginès i Fabrellas, Anna. "Analítica de personas y discriminación", op. Cit.; y Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women", op. Cit.; véase también el informe "La discriminación algorítmica en España: límites y potencial del marco legal", elaborado en 2022 por Digital Future Society y en el que se recogen otros tantos. Disponible en: https://digitalfuturesociety.com/app/uploads/2022/09/Discriminacion_algoritmica_Espana_marco_legal.pdf

9), con especial atención a colectivos vulnerables y grupos protegidos, tal y como se contempla en el reglamento europeo de IA[6].

Para que esta intervención temprana trascienda el plano consultivo y adquiera fuerza normativa, la negociación colectiva debe erigirse en un instrumento capaz de incorporar límites materiales y garantías jurídicas al algoritmo, traduciendo las observaciones de la representación laboral en cláusulas convencionales que verdaderamente condicionen el diseño y el uso del sistema. Estas cláusulas podrían quedar estructuradas en torno a tres ejes de protección específicamente relevantes en el contexto laboral:

El primero sería la prohibición de sustitución encubierta del poder disciplinario mediante automatización. En efecto, existe un riesgo particular en el hecho de que los algoritmos se utilicen para tomar decisiones sancionadoras o para determinar de modo automático las consecuencias disciplinarias de las desviaciones de la persona trabajadora respecto a estándares fijados por el propio sistema. La sustitución algorítmica del juicio humano en una esfera especialmente sensible como es la del Derecho del Trabajo, donde el ordenamiento exige garantías reforzadas (audiencia previa, motivación clara de la sanción, proporcionalidad entre incumplimiento y consecuencia, y prohibición de discriminación), podría tener consecuencias trascendentales como es la eliminación o reducción drástica de la posibilidad de contradicción, las motivaciones opacas que impiden la impugnación efectiva, la aplicación de reglas uniformes sin ponderar circunstancias individuales (agravantes o atenuantes), la intensificación del control mediante consecuencias negativas encubiertas (como puede ser la asignación de peores turnos, la reducción de oportunidades de trabajo o bonus, o la rebaja de puntuaciones internas) que dificultan su identificación (esto es, la inexistencia de un acto formal de sanción obstaculiza la posibilidad de conocer que se está siendo penalizado y, por tanto, también su impugnación), o la inversión de la carga de la prueba al conferir presunción de veracidad al dato algorítmico (lo que la persona trabajadora debe rebatir sin acceso a la lógica del sistema). Por todo lo anterior, la cláusula convencional debe establecer que el algoritmo no puede adoptar por sí mismo decisiones que supongan sanciones, despidos o cambios sustanciales de las condiciones de trabajo; antes bien, cualquier decisión disciplinaria debe transitar por un cauce humano dotado de garantías mínimas, incluyendo audiencia de la persona afectada y motivación clara de la decisión, accesible más allá de la mera referencia a recomendaciones del sistema automatizado.

El segundo eje atañe a los límites sobre monitorización y control intensivo. La negociación colectiva puede establecer exclusiones explícitas respecto a tipos de datos que no pueden integrar el modelo. Este sería el caso de los datos de geolocalización

6. 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 (DOUE núm. 1689, de 12 de julio de 2024).

fuera de la jornada de trabajo; la actividad en redes sociales personales; los datos biométricos que vayan más allá de lo estrictamente necesario para la finalidad alegada; o de la información que revele aspectos de la vida íntima de la persona trabajadora.

Igualmente importante es que se fijen criterios de proporcionalidad para el seguimiento del rendimiento. Así, aunque se admita el uso de ciertos datos de productividad o calidad, su incorporación al algoritmo debe estar sujeta a límites de frecuencia temporal, a exclusiones respecto a momentos especialmente sensibles de la jornada, y a cálculos de impacto que verifiquen que no generan una intensificación patológica del control[7].

El tercer eje debe ir referido a la garantía del derecho a la desconexión y a la vida privada en la configuración del algoritmo. Específicamente, la cláusula convencional debe vedar que el sistema incorpore datos generados fuera del tiempo de trabajo (salvo excepciones que requieran consentimiento informado y justificación estricta), a la vez que debe incluir salvaguardas que impidan que el algoritmo penalice de facto el ejercicio legítimo del derecho a desconexión. Un ejemplo paradigmático de lo contrario sería un sistema de asignación de turnos o de oportunidades que discrimine a trabajadores y/o trabajadoras que no responden a mensajes o solicitudes fuera de su horario contractual, puesto que eso equivaldría a convertir la desconexión en un coste profesional, vaciando así de contenido ese derecho.

Todas estas cláusulas convencionales deben acompañarse además de la exigencia de que la empresa realice y ponga a disposición de la representación laboral evaluaciones de impacto previas, es decir, estudios realizados antes de la implantación del sistema que valoren cómo el diseño previsto afectará a la igualdad de trato entre grupos, a la salud y el bienestar laboral (de manera particular en lo referente a riesgos psicosociales derivados del control automatizado), y al ejercicio efectivo de derechos colectivos. Estas evaluaciones no deben ser meramente técnicas, por otro lado, sino que deben integrarse en el diálogo jurídico-laboral y ser comunicadas a la representación legal como documentación sustancial de la fase de diseño, permitiendo así que los y las representantes verifiquen que sus observaciones han sido integradas o, en su defecto, que conozcan las razones por las cuales la empresa decide proceder de un modo distinto.

En síntesis, la fase de diseño deja de ser, en este marco de gobernanza algorítmica laboral, un espacio técnico cerrado dominado por especialistas en ingeniería de datos e inteligencia artificial para transformarse en un terreno genuinamente de deliberación jurídico-laboral, donde se fijan los contornos constitucionales de lo que el algoritmo puede y no puede hacer en el seno de la empresa, donde se establecen desde el origen las protecciones contra la discriminación y el control desmesurado, y donde

7. El propio artículo 9 del Reglamento europeo de IA alude a esta granularidad temporal de los datos en la identificación de riesgos.

se garantiza que los colectivos más expuestos no vean agravada su posición de vulnerabilidad por un diseño de IA que reproduce o amplifica las desigualdades ya existentes. Esta redefinición de la fase de diseño como momento de gobernanza colectiva constituye una de las contribuciones más relevantes de la negociación colectiva a la construcción de una IA laboral verdaderamente equitativa.

3. Fase de entrenamiento y pruebas. De la teoría a la práctica

Como se apuntaba al inicio, en esta etapa, la gobernanza algorítmica laboral se desplaza desde el plano de las decisiones de diseño del sistema al de la verificación empírica de las consecuencias de su aplicación en el ámbito del trabajo. Esto es, el sistema, alimentado con datos históricos, empieza a generar predicciones y recomendaciones de decisión, lo que constituye una segunda oportunidad de intervención negocial, pues es en este momento cuando la de RLT debe poder comprobar si el modelo reproduce desigualdades previas o genera patrones de decisión incompatibles con la igualdad u otros derechos esenciales.

Durante esta fase de entrenamiento, al algoritmo se le proporcionan datos (por ejemplo, perfiles de personas candidatas seleccionadas y no seleccionadas, evaluaciones de desempeño históricas, decisiones de promoción previas, etc.). Éste los analiza y extrae patrones (por ejemplo, cuando un/a candidato/a tiene "X" años de experiencia e "Y" tipo de formación, suele ser seleccionado/a. Cuando tiene estas otras características, suele ser rechazado/a). El modelo desarrolla así miles de reglas que le permiten, ante una nueva persona candidata, estimar si debiera ser seleccionada o no. Y aquí comienza la clave, pues esas reglas no tienen por qué ser simples ni transparentes. Puede tratarse de combinaciones complejas de centenares o miles de variables, con relaciones no lineales entre ellas. Por ejemplo, el algoritmo podría aprender que una interrupción laboral de 2 años penaliza cuando la persona tiene entre 30 y 40 años, pero no cuando tiene menos de 25 años; y además que esa penalización se multiplica si la persona es mujer y vive en determinada zona geográfica.

En las técnicas de IA actuales, el aprendizaje del propio modelo no se puede predecir a priori (no olvidemos que se trata de un sistema de aprendizaje automático). Es lo que se conoce como caja negra (*black box*) [8]. El modelo aprende de los datos históricos y arroja una decisión, pero el mecanismo interno o la secuencia seguida para llegar a ella permanece opaca (siguiendo con el ejemplo anterior, entra un/a candidato/a, sale una recomendación de decisión). Si los datos que se han introducido para

8. Sobre este fenómeno y la explicabilidad en la IA hay mucha literatura. Sirva de ejemplo, el documento del Supervisor Europeo de Protección de Datos, "TechDispatch: Explainable Artificial Intelligence", de noviembre 2023. Disponible en: https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf

entrenar no son neutrales, el modelo reproducirá e, incluso, ampliará los sesgos existentes. Por ejemplo, si una empresa ha contratado históricamente más hombres que mujeres para ciertos puestos, el modelo puede aprender que los hombres son mejores candidatos para esos puestos. De la misma manera que si la empresa ha promocionado principalmente a personas sin interrupciones en la carrera, es probable que penalice a quienes tuvieron alguna baja por maternidad o enfermedad. Y es que el modelo lo que hace es buscar patrones estadísticos que correlacionen con lo que queremos saber. Si históricamente las mujeres fueron menos seleccionadas, el modelo detectará que ser mujer es un predictor estadístico de "no ser seleccionado", y usará esa correlación para rechazar a futuras candidatas, incluso si sus habilidades reales son equivalentes a las de candidatos hombres seleccionados en el pasado. Más aún, el algoritmo puede detectar que variables como ciertos tipos de universidades, códigos postales, o interrupciones en la carrera correlacionan con género, y usa esas correlaciones para amplificar la discriminación de forma indirecta.

En consecuencia, los sesgos se generan durante el entrenamiento, no después. Si dejamos que el algoritmo entrene con datos históricos sesgados sin intervención, las decisiones discriminatorias ya estarán aprendidas en el modelo.

Lo paradójico es que el algoritmo parece objetivo (está basado en datos; es un proceso matemático; y no tiene prejuicios conscientes como un humano). Pero precisamente porque es posible que los datos históricos que usa contengan sesgos humanos y porque esos sesgos van a estar codificados en correlaciones estadísticas complejas resulta más difícil identificar y rebatir la discriminación que si la cometiera un humano de forma consciente. Por ello, la fase de entrenamiento constituye un momento crítico para introducir controles que no existían en la fase de diseño.

Si en la primera etapa la representación de las personas trabajadoras podía opinar sobre intenciones, criterios generales y límites normativos, en esta fase debe ser capaz de verificar el comportamiento real del modelo. Concretamente, debería disponer, en términos claros, de información sobre:

a) Tasas de error diferenciadas por grupos protegidos. Si el modelo comete más errores o produce decisiones menos favorables para mujeres que para hombres, para trabajadores mayores que para más jóvenes, o para personas con discapacidad que, para personas sin discapacidad, por ejemplo.
b) Distribución de puntuaciones asignadas por el modelo. No sólo promedios, sino información sobre cómo se distribuyen las puntuaciones entre grupos. Por ejemplo, si el modelo tiende a asignar puntuaciones más bajas de forma sistemática a candidatos/as de determinados orígenes, o si amplifica pequeñas diferencias iniciales en grupos determinados.
c) Relevancia de las variables en las decisiones del modelo. Es decir, qué factores ha aprendido a considerar el algoritmo como más relevantes. Si emerge

que una variable aparentemente neutral (como puede ser la localización residencial, el tipo de institución educativa, o la duración de interrupciones en la carrera) se ha convertido en un proxy[9] altamente influyente para características protegidas, se revelaría una forma de discriminación indirecta potencial que debe ser corregida antes de la implantación generalizada del modelo.

d) Esta información debe presentarse a los y las representantes en términos comprensibles, no como parámetros técnicos indescifrables. Esto podría facilitarse en forma de respuestas a preguntas simples: ¿el modelo tiende a rechazar a más mujeres que a hombres en proporción relativa?, ¿asigna puntuaciones más bajas a personas mayores de 55 años?, ¿hay una variable correlacionada con género que sea particularmente determinante en las decisiones? Sólo con una información accesible puede la representación laboral formar un juicio sobre si el modelo es aceptable o requiere correcciones. En otras palabras, el valor añadido para la RLT no está en supervisar la técnica de programación, sino en contrastar, caso a caso o mediante patrones, si las decisiones algorítmicas de prueba son coherentes con los criterios convencionales y legales aplicables, y si generan efectos sistemáticamente adversos para determinados colectivos.

Una vez disponible la información sobre el comportamiento del modelo en laboratorio, la segunda intervención de la representación toma forma mediante su participación en pruebas piloto controladas. Esto es, el modelo no debe pasar directamente de las pruebas técnicas a la aplicación generalizada en toda la empresa o en toda la plantilla, sino que debería desplegarse de forma limitada (en un centro de trabajo específico, para una categoría profesional determinada o para un tipo de decisión acotado -como puede ser la evaluación de desempeño, pero no el cálculo de incentivos-) y durante un período de tiempo definido.

La finalidad de esta participación es doble. En primer lugar, reducir la brecha entre las predicciones del modelo en el laboratorio y su comportamiento en la práctica real. Esta participación permite a la RLT observar si aparecen patrones inesperados, comportamientos indeseables o incoherencias con los criterios laborales ya aplicados; lo que, a su vez, posibilitaría analizar por qué el modelo ha adoptado esas decisiones mediante *técnicas de interpretabilidad algorítmica accesibles*[10]. En segundo lugar, y

9. Remito nuevamente a lo ya reseñado en la nota al pie 4.

10. Existen técnicas de interpretabilidad algorítmica accesibles, como LIME, SHAP (aunque ambas precisan de un experto en IA para llevar a cabo la traducción entre el lenguaje técnico y las conclusiones comprensibles) o los árboles de decisión (más accesibles de forma inherente, especialmente si son pequeños), que permiten traducir la lógica compleja del modelo a explicaciones comprensibles para no expertos, identificando qué factores han pesado más en una decisión concreta. Véase, *La IA explicable*, IBM, Think, 2025. Disponible en: https://www.ibm.com/es-es/think/topics/explainable-ai

quizás más importante, dicha participación protege a las personas (trabajadoras y/o candidatas) frente a posibles daños de un modelo aún experimental. Lo que se conseguiría mediante un doble compromiso empresarial expreso. Por un lado, la empresa debe asumir que los resultados de las decisiones algorítmicas durante esta fase piloto no tendrán consecuencias irreversibles ni sancionadoras, garantizándose así que las personas no van a quedar expuestas a riesgos innecesarios. Por otro, resulta igualmente esencial que exista un compromiso de rectificación. La empresa ha de asumir el necesario ajuste del modelo, e incluso descartar su implantación, si se constata que sus efectos son incompatibles con los principios de igualdad y no discriminación. Sólo así, lo que en principio se plantea como una prueba unilateral, se transforma en un espacio seguro de experimentación conjunta.

4. Evaluación de impacto específica. Una fase intermedia

Antes de que el sistema se generalice más allá del piloto, la gobernanza algorítmica laboral que aquí propongo exige un paso más, que bien podría constituir el último paso de la etapa anterior. Me estoy refiriendo a la realización una evaluación de impacto específicamente diseñada para el contexto laboral y la protección de derechos.

Esta evaluación es distinta de la evaluación de impacto sobre los derechos fundamentales exigida por el artículo 27 del reglamento IA europeo, aunque complementaria. Motivo por el que he preferido dedicarle un apartado específico en este planteamiento.

En base a sus caracteres principales podrían sintetizarse ambas a través de la siguiente comparación gráfica.

Tabla 1
Comparación gráfica

	EVALUACIÓN DE IMPACTO (ART 27)	EVALUACIÓN DE IMPACTO LABORAL (PROPUESTA)
QUIÉN LA EXIGE	Legislador europeo	La negociación colectiva
A QUIÉN SE APLICA	Todos los proveedores de sistemas de IA de alto riesgo (incluyendo laborales)	Específicamente a algoritmos usados en decisiones sobre empleo
QUÉ ANALIZA	Riesgos genéricos para derechos fundamentales de forma abstracta y amplia	Impactos reales y concretos en el empleo y la vida laboral de trabajadores/as específicos/as.

	EVALUACIÓN DE IMPACTO (ART 27)	EVALUACIÓN DE IMPACTO LABORAL (PROPUESTA)
CÓMO	Evaluación de riesgo potencial o teórico	Evaluación de impacto empírico y medible
RESULTADO	Documento normativo que certifica que se han considerado riesgos, pero no necesariamente cuantifica ni mide impactos reales sobre trabajadores/as específicos/as.	Informe con datos reales, estadísticas fiables e impactos medidos, que permite a la RLT tomar decisiones informadas sobre si el algoritmo es aceptable.

Fuente: elaboración propia

Mientras que con la evaluación del artículo 27 lo que se lleva a cabo es un control de conformidad normativa, orientado a que las autoridades supervisoras puedan intervenir si hay violaciones graves, con la que aquí se propone se ejecutaría un control de equidad práctica. En otras palabras, la primera es demasiado genérica para el contexto laboral.

Para la RLT, lo relevante es conocer si esos riesgos de discriminación se han materializado realmente en los datos propios, en qué medida y/o a cuántas personas trabajadoras afecta. Esto es, si el modelo funciona adecuadamente en el contexto propio y específico, pues eso es lo que verdaderamente va a orientar su labor para negociar condiciones de uso o mejoras.

Imaginemos que en nuestra empresa se plantea la implantación de un algoritmo de selección. La evaluación de impacto europea arrojaría como resultado algo parecido a esto: "El algoritmo utiliza datos históricos de contratación. Existe un riesgo potencial de que reproduzca sesgos de género presente en esos datos". Frente a lo cual, la empresa implementaría la supervisión humana para mitigar el riesgo. Como resultado de la evaluación se certificaría que hay riesgos identificados, pero que se articulan medidas de mitigación, por lo que formalmente cumpliría con la normativa europea.

Si sometemos el mismo supuesto a la evaluación de impacto laboral propuesta, el resultado se parecería más a lo siguiente: "Durante 6 meses de piloto, el algoritmo procesó 500 candidaturas. Rechazó al 78% de candidatas mujeres frente al 42% de candidatos hombres. Las variables con mayor importancia fueron: universidades de procedencia (proxy de género) y duración de interrupciones en carrera (proxy de maternidad). En el análisis de impacto se detectó que el 65% de rechazos de mujeres se debían a estos dos factores. Las revisiones humanas sólo corrigieron el 15% de los casos discriminatorios. Se recomienda: exclusión de estas variables, aumento del umbral de revisión humana, o no usar el sistema para la selección inicial". Como puede

apreciarse, no sólo cuantifica exactamente qué está pasando, sino que permite identificar el mecanismo de discriminación y proponer soluciones concretas.

Como indicaba al inicio de este apartado ambas evaluaciones no son excluyentes; de hecho, se complementan de manera necesaria. Sin la evaluación de la norma europea, el algoritmo podría violar derechos fundamentales sin que la empresa lo sepa anticipadamente. Pero a la vez, sin una evaluación de impacto laboral, sería posible la paradoja de que la empresa cumpliría formalmente con el mandato legal, pero no sabría si el algoritmo está efectivamente discriminando a sus trabajadores/as en la práctica.

En suma, esta evaluación específica del contexto laboral, cuantificada empíricamente e inmediatamente operativa para la negociación colectiva, permite a la RLT disponer de datos concretos sobre cómo el algoritmo afecta realmente a personas trabajadoras y candidatas específicas para poder negociar condiciones de uso, medidas de mitigación o incluso rechazar el sistema si los impactos son incompatibles con la igualdad y los derechos laborales.

5. Campo para la negociación durante la implantación y el uso efectivo del modelo

La transición del modelo desde un sistema experimental (fase piloto) hacia su implantación generalizada en la empresa marca un punto de inflexión en la gobernanza algorítmica laboral. No se trata simplemente de activar un sistema ya validado, sino de establecer un marco de reglas, garantías y mecanismos de control que protejan los derechos de las personas trabajadoras durante todo el tiempo que el algoritmo influya en decisiones laborales. Este marco resulta idóneo para la intervención de la negociación colectiva. A ella debe corresponder su construcción en torno a dos momentos clave: la definición inicial de garantías (momento constituyente) y la vigilancia continuada durante el funcionamiento (momento de control permanente).

5.1. Definición del marco de garantías: el momento constituyente de reglas y límites

Antes de que el algoritmo empiece a operar de forma generalizada es imprescindible que la RLT negocie y pacte por escrito un conjunto de garantías que actúen como límites vinculantes al uso del sistema. Estas garantías deben materializarse en cláusulas de convenio colectivo, acuerdos de empresa o, en su defecto, en compromisos formales documentados que tengan un estatus equiparable (como pudiera ser el caso, por ejemplo, de un Protocolo de uso de algoritmos y derechos digitales firmado por la empresa y el Comité de Empresa).

Con idea de establecer un escudo defensivo mínimo para proteger los derechos laborales frente a la automatización se proponen a continuación cuatro garantías, basadas

en los mismos pilares estructurales de cualquier sistema de garantías laborales: *Saber* (Información); *Influir* (Consulta), *Controlar* (Supervisión); y *Defenderse* (Impugnación).

Desde esa lógica se proponen como básicas las siguientes cláusulas:

a) Información previa y comunicación estructurada a la plantilla. La plantilla afectada debe recibir, antes de la implantación generalizada, información clara, comprensible y suficiente sobre qué significa la introducción del algoritmo en su vida laboral. Esta información no debe ser un comunicado técnico incomprehensible, sino una explicación accesible que responda a preguntas concretas: ¿en qué decisiones interviene el sistema (selección, turnos, evaluación, incentivos) ?, ¿qué datos personales utiliza?, ¿quién puede revisar o impugnar una decisión?, ¿cuáles son mis derechos como trabajador/a? La información debe transmitirse además a través de canales múltiples (reuniones informativas, documentos escritos, plataformas digitales, etc.) y debe estar disponible en diferentes idiomas si la plantilla es diversa. La RLT debe poder verificar además que la información es completa, accesible y veraz, de la misma manera que debe tener derecho a participar en su elaboración y difusión. Esto no debe ser entendido como un mero trámite administrativo, sino como una garantía de que los trabajadores y las trabajadoras no se enfrentan a un sistema opaco.

b) Consulta preceptiva sobre los términos de la implantación en torno a los siguientes elementos operativos:

   — Calendario y ritmo de aplicación: ¿se implanta el sistema en toda la empresa simultáneamente o en fases (por departamentos, por tipo de decisión, etc.) ?, ¿cuánto tiempo de transición hay (es decir, durante el que algoritmo va a funcionar en paralelo con procedimientos humanos antes de ser el único método de decisión) ?, ¿desde qué momento empezará a influir en sus decisiones?
   — Colectivos y decisiones afectadas: ¿a qué categorías de personas trabajadoras se aplica el algoritmo?, ¿hay colectivos que quedan expresamente excluidos (por ejemplo, personas con discapacidad)?
   — Criterios de priorización en caso de que el algoritmo no pueda procesar todos los datos a la vez. ¿Cuál es el criterio para decidir quién accede primero a las oportunidades que se puedan platear? Estos criterios deben ser explícitos y pactados. No puede tratarse de una decisión unilateral técnica (por ejemplo, porque quien lo programara decidió procesarlos alfabéticamente). ¿Es justo y no discriminatorio?, ¿es transparente? Las personas candidatas o trabajadoras

afectadas deben saber que existe ese criterio de priorización. Se trata de elemento que suele pasarse por alto porque parece de carácter técnico, pero en realidad es de justicia laboral, pues en definitiva va a determinar quién tiene oportunidades y quién no.

— Finalmente, aunque excede el estricto marco de la gobernanza técnica del algoritmo, la consulta preceptiva no puede ignorar los posibles impactos organizativos. De manera que, si la automatización implicara cambios sustanciales en funciones o puestos de trabajo, debería activarse en paralelo la negociación de medidas de mantenimiento del empleo y recualificación profesional.

Esta consulta preceptiva no es meramente formal. La RLT debe tener la capacidad de bloquear o retrasar un despliegue que considere injusto, excesivamente rápido, o insuficientemente acompañado.

c) Pactos vinculantes sobre supervisión humana cualificada. Se trata de uno de los elementos más críticos de la gobernanza propuesta. Y es que la supervisión humana no puede ser un concepto abstracto. Debe concretarse en procedimientos observables y verificables, por lo que la cláusula convencional debe especificar:

— Quién supervisa: ¿qué personas, departamento o comisión tienen la responsabilidad de revisar?, ¿tienen formación específica?, ¿tienen independencia suficiente frente a presiones de eficiencia o costes?
— Qué tipo de decisiones requieren supervisión obligatoria. Esto implica la permisibilidad de decisiones automáticas de bajo impacto (como pudieran ser las recomendaciones de formación), siempre que haya mecanismos de impugnación posteriores.
— Plazos de revisión que han de ser razonables para que la revisión sea real y no un trámite vacío.
— Garantías de motivación: la persona que revisa una decisión algorítmica debe poder acceder a una explicación comprensible del porqué el sistema llegó a esa decisión.
— Prioridad de la decisión revisada: si tras la supervisión humana se detecta que la decisión algorítmica era injusta o discriminatoria, la decisión revisada por humanos debe tener prioridad absoluta.
— Estos parámetros transforman la supervisión humana de una promesa vaga (“habrá supervisión”) a un procedimiento controlable.

d) Procedimientos de impugnación y recursos claros y accesibles para que las personas trabajadoras y candidatas puedan ejercer realmente sus derechos. Esto incluye, en particular, lo siguientes derechos:

— Derecho a ser informado. Cuando una decisión que afecta negativamente a la persona (como puede ser un rechazo de candidatura o una penalización) haya sido recomendada o adoptada por el algoritmo, ésta debe ser informada expresamente.
— Derecho de explicación. No me refiero aquí a que la empresa tenga que entregar el código fuente o la fórmula matemática, sino a que la persona pueda solicitar la motivación laboral de la decisión concreta. Para poder dar esta respuesta, la empresa está obligada a usar herramientas de interpretabilidad que permitan identificar qué variables o factores fueron determinantes para la decisión. Si la empresa usa un algoritmo tan opaco que no puede identificar ni siquiera qué variables pesaron más, entonces ese algoritmo no debería usarse para decisiones laborales, porque incumpliría el deber básico de motivación de los actos empresariales[11].
  La empresa debe además proporcionar esta explicación en un plazo breve (por ejemplo 5-10 días).
— Derecho a impugnar la decisión y a solicitar su revisión. La persona debe poder presentar una reclamación formal sobre la base de que la decisión es injusta, discriminatoria, o está basada en datos erróneos e instar, en consecuencia, su revisión por una persona o comisión distinta de quien tomó la decisión original. De hecho, debería ser posible solicitar que la revisión fuera llevada a cabo exclusivamente por un/a humano/a, sin que haya vinculación a la recomendación del algoritmo.
— Plazos claros (por ejemplo, máximo 15 días para responder a una impugnación) y establecimiento de un mecanismo de escalada para que, si la empresa rechaza la impugnación, la RLT o un tercero imparcial puedan intervenir con garantías.
— Protección frente a represalias. Cualquier persona trabajadora o candidata que impugne una decisión algorítmica debe estar protegida frente a posibles represalias, cambios en condiciones de trabajo, o estigmatización.

11. Lo que encuentra sustento igualmente en el art. 22 y el considerando 71 del RGPD, que reconocen el derecho a obtener una explicación significativa de la lógica de una decisión automatizada.

5.2. Control permanente o función de vigilancia continuada

Una vez que el sistema está en funcionamiento, la gobernanza algorítmica laboral no termina. Muy al contrario, comienza la fase más activa para la RLT: la supervisión continua de cómo el sistema se comporta en la realidad. Esto es, las reglas y garantías pactadas en el momento constituyente deben verificarse, mantenerse, y ajustarse si es necesario.

En esta labor, la RLT debe disponer de acceso regular a datos sobre cómo el algoritmo está funcionando en la práctica. Por ejemplo, tasas de error identificado cometido por el algoritmo; tasa de rechazo, penalización o exclusión diferenciada según parámetros (sexo, edad, tipo de contrato, nacionalidad, discapacidad, etc.); total de decisiones que han sido impugnadas; tasa de éxito de las apelaciones (incluyendo patrones si hubiera determinados tipos de decisión que son apeladas con mayor frecuencia); porcentaje de decisiones que han sido sometidas a revisión humana y, de ellas, proporción en la que se ha rechazado o modificado la recomendación del algoritmo; posibles actualizaciones implementadas en el modelo, con indicación de si se informó o no a la RLT y se realizó una nueva evaluación de impacto. Se trata, en suma, de información de carácter genérico que va a permitir vigilar posibles tendencias en el modelo.

Por otro lado, esta información debe ser sistemática y periódica, no discrecional ni controlada unilateralmente por la empresa. Igualmente debe presentarse en un formato claro y visual (gráficos, tablas, resúmenes ejecutivos), a ser posible, y no tratarse de un mero volcado bruto de datos o informes técnicos ininteligibles. Con este tipo de acceso a esa información lo que se busca es transparencia (es lo que va a permitir conocer de manera ágil cómo va el sistema) y facilitar el control pasivo.

Ahora bien, no sería operativo (ni seguro por protección de datos) volcar toda la información detallada a toda la RLT con mayor o menor periodicidad. Por ello, se propone un sistema escalonado: transparencia general para toda la RLT, capacidad de conocimiento detallado y específico para un órgano especializado: la comisión paritaria de seguimiento y auditoría, con participación equilibrada de empresa y RLT. Entre sus funciones se contempla el deber de reunirse regularmente para revisar el funcionamiento del algoritmo; el acceso a datos e información detallada (posibilidad de solicitar explicaciones ampliadas, acceder a casos específicos de decisiones controvertidas, y/o examinar la lógica del modelo con el apoyo de personal experto técnico, si es necesario); la identificación de patrones anómalos o incoherentes con los criterios laborales aplicados previamente; así como la capacidad de proponer correcciones si identifica un sesgo o un problema, o incluso, la suspensión del sistema si el daño es grave (función que debería ser reforzada con la capacidad de ser vinculante para la empresa); y finalmente la consulta preceptiva, vinculante y previa a cualquier cambio que pretenda introducir la empresa en el sistema.

La introducción de este órgano especializado transforma la gobernanza de un sistema reactivo, que esperara que haya un problema, a uno proactivo con capacidad para detectar problemas antes de que causen daños masivos.

En esta fase resulta esencial contar además con procedimientos de impugnación de las decisiones algorítmicas, más ágiles y accesibles que en otras decisiones empresariales (por un motivo simple: el riesgo de decisiones masivamente injustas o discriminatorias es más alto con algoritmos que con decisiones humanas puntuales). Entre estas medidas específicas podrían incluirse:

— Canales simplificados de impugnación: acceso directo a la comisión paritaria o a la figura de mediación independiente, sin necesidad de pasar por todos los trámites internos que puedan existir en el cauce general de reclamación previsto en la empresa.
— Inversión de la carga de la prueba. Especialmente cuando hay indicios de sesgo sistemático, la empresa debe probar que la decisión del algoritmo fue justa, en lugar de que el trabajador deba probar que fue injusta.
— Derecho a prueba pericial externa, ante una duda razonable basada en datos. Aspecto que seguro suscita controversia. No me refiero aquí a un derecho automático de auditoría a demanda de la RLT, sino a aquellos casos en los que existe una duda proporcionada o acorde a los datos existentes (por ejemplo, estadísticas que muestran una disparidad significativa).
Imaginemos que el tema en discusión es que el algoritmo discrimina a las mujeres. La empresa dirá "mis técnicos dicen que no". La RLT dirá "nosotros creemos que sí". Se produce una situación de asimetría técnica insalvable. La empresa tiene los datos, el código y los/as ingenieros/as. La RLT, por lo general, no tiene nada de eso. Sin un tercero independiente (un/a perito), la RLT está ciega y la disputa entra en punto muerto o acaba en juicio (donde, paradójicamente, el juez acabará pidiendo un/a perito).
Podría plantearse algo parecido a lo siguiente:

> Mecanismo de desbloqueo técnico: en caso de discrepancia fundada sobre el funcionamiento del algoritmo, que no pueda resolverse en la comisión paritaria, las partes podrán acordar el nombramiento de una persona auditora técnica independiente, sometida a confidencialidad.

Quedaría pendiente, no obstante, quién ha de asumir su coste. Cuestión nada baladí, dado el elevado importe de las auditorías algorítmicas[12]. Si bien,

12. La complejidad multidisciplinar exigida por guías oficiales como la de la AEPD implica la intervención de perfiles técnicos y jurídicos de alto coste (AEPD - Guía de Auditoría de Tratamientos que incluyan IA (2021/2024). Disponible en: https://www.aepd.es/guias/

una solución neutral podría ser vincular el pago al resultado del dictamen; de tal manera que, s*i la auditoría confirma que había un error o sesgo,* el coste lo asume la empresa. Por el contrario, si el algoritmo estaba bien, el coste habría de abonarlo quien pidió la auditoría.

— Retroactividad de los efectos si se detecta que el algoritmo ha estado discriminando de forma sistemática durante un período. En tal caso deben revisarse todas las decisiones del período afectado y repararse los daños ocasionados (indemnizaciones, restitución de oportunidades, etc.).

Siguiendo con la línea planteada al inicio, durante la vida útil del algoritmo, también surgirán situaciones que requieran cambios: nuevos datos, cambios en la legislación, en la organización del trabajo, identificación de nuevos sesgos, etc. Situaciones frente a las que la negociación colectiva debe prever, además:

— Un protocolo de actualización que contemple la consulta preceptiva a la RLT antes de cualquier cambio significativo del modelo, una nueva evaluación de impacto y el acuerdo necesario sobre si el cambio es aceptable.
— Evaluaciones de impacto periódicas para verificar que el modelo sigue siendo equitativo y no ha desarrollado nuevos sesgos a lo largo del tiempo.
— El derecho a proponer la desactivación, si la RLT considera que el algoritmo ya no es compatible con los estándares de igualdad y protección de derechos. Si la propuesta es justificada (basada en datos), la empresa debe considerarla seriamente y, en su caso, negociar un plan de transición hacia alternativas.

Esta estructura permite que la negociación colectiva no solo influya en el diseño del algoritmo, sino que mantenga una presencia activa durante toda la vida útil del sistema, disponiendo de información, cuestionando patrones problemáticos, proponiendo correcciones, y, en último término, siendo capaz de detener un sistema que se haya vuelto injusto. Solo así la gobernanza algorítmica laboral deja de ser un concepto abstracto para convertirse en una práctica de protección real y continuada de derechos laborales.

6. Fase de desmantelamiento o sustitución. El refuerzo negocial para la continuidad de derechos y garantías

La vida útil de un sistema algorítmico no es indefinida. Con el tiempo, es posible que la empresa decida retirar el algoritmo, sustituirlo por una versión mejorada, migrarlo

requisitos-auditorias-tratamientos-incluyan-ia.pdf). Igualmente, programas públicos de fomento digital como Kit Consulting (programa de ayudas del Gobierno de España impulsado por Red.es) sitúan una consultoría básica de IA por encima de los 6.000 euros en el ámbito de las pymes, lo que indica que una auditoría profunda en una gran empresa tendrá costes significativamente superiores.

a una nueva plataforma más integral, o volver a decisiones puramente humanas. Este momento de transición es crítico desde la perspectiva de la gobernanza algorítmica laboral porque concentra dos riesgos de manera simultánea: la posible desaparición de evidencias sobre decisiones pasadas y los impactos organizativos del cambio tecnológico, que pueden afectar seriamente a derechos adquiridos por las personas trabajadoras.

En la senda de esta visión procesual (y no estática), en la que la gobernanza acompaña al algoritmo desde su concepción hasta su retirada resulta imprescindible volver a proyectar *ex ante* el papel garantista de la RLT en esta fase crítica para la continuidad de derechos y garantías.

En este sentido, cuando la empresa decide retirar, sustituir o modificar sustancialmente el algoritmo, debe informar de forma específica y con la antelación suficiente a la RLT y a la Comisión Paritaria sobre la decisión adoptada. Esta información complementa y cierra el ciclo de información periódica previsto durante el funcionamiento ordinario, y debe especificar las circunstancias que sustentan la decisión empresarial de retirarlo, así como el plan de ejecución.

Las razones del desmantelamiento o sustitución del sistema pueden ser reveladoras. Así si el sistema se quita porque existe evidencia de discriminación, hay una admisión empresarial de facto que puede ser utilizada en reclamaciones de personas afectadas en el pasado. Del mismo modo que si se retira por razones económicas (el nuevo sistema cuesta menos, por ejemplo), la RLT puede anticipar posibles reestructuraciones.

Mayor relevancia tiene, no obstante, conocer qué sucede con los datos personales de los/as trabajadores/as almacenados por el algoritmo. ¿Se transfieren al nuevo sistema (con riesgo de perpetuar sesgos o errores previos) ?; ¿se conservan para una auditoría posterior?; ¿se eliminan conforme al derecho al olvido del RGPD? Y, si es así ¿con qué cronograma? Esta información es esencial porque la preservación de datos históricos es la única forma de que aquellas personas que sufrieron decisiones injustas puedan, posteriormente, demostrar que fueron discriminadas. En otras palabras, uno de los mayores riesgos del desmantelamiento de un sistema algorítmico es la desaparición de pruebas. Pensemos en el caso de un trabajador que fue rechazado por un algoritmo hace ocho meses y decide reclamar seis meses después. La empresa puede alegar que "el sistema ya no existe y que no es posible recuperar los datos". Sin esos datos, el trabajador queda jurídicamente indefenso al no poder probar cómo el algoritmo llegó a la decisión que le perjudicó, y la empresa no puede refutarlo. Por ello, la gobernanza algorítmica laboral debe incluir explícitamente un derecho a preservación de evidencias que implique:

— La obligación de la empresa de conservar los registros de todas las decisiones tomadas o recomendadas por el algoritmo, incluidos los datos de entrada

utilizados y los parámetros que pesaron en cada decisión durante un período suficiente (mínimo el equivalente al plazo de prescripción de derechos laborales).
— La accesibilidad a esos registros para la defensa de derechos, lo que habrá de compaginarse con garantías de confidencialidad debidas frente al posible acceso de otras personas afectadas.
— El deber empresarial de conservar documentación sobre cómo funcionaba el algoritmo (qué variables utilizaba, qué pesos tenían, cómo se entrenó, qué pruebas de sesgo se realizaron, etc.). Si el algoritmo desaparece sin esa documentación, será imposible saber, años después, si fue o no discriminatorio.

Sin esta preservación, el desmantelamiento de un sistema algorítmico merma considerablemente la capacidad de defensa de los derechos lesionados mientras el sistema funcionaba.

Distinto es el caso de que la empresa implantara un algoritmo nuevo o volviera a métodos humanos. Este cambio puede traer consecuencias laborales sustanciales frente a las que la RLT debe poder negociar protecciones durante la transición. Como, por ejemplo:

— La garantía de continuidad de derechos adquiridos. Si una persona trabajadora acumula determinados derechos conforme a cómo el sistema A evaluaba su desempeño (antigüedad, puntuación, acceso a oportunidades), la migración al algoritmo B o a criterios humanos no puede retrotraer esos derechos. Es decir, si bajo el algoritmo A, su labor fue evaluada como "rendimiento alto" y, bajo el nuevo sistema debería haber sido "rendimiento medio", la empresa no puede pretender que el cambio se aplique retroactivamente (recortando bonificaciones o incentivos pasados).
— Formación y apoyo en la transición en caso de que el cambio de sistema requiera que la plantilla y las personas encargadas de la supervisión aprendan nuevas formas de trabajar.
— Protección frente a ajustes de plantilla asociados al cambio. La negociación colectiva debe establecer que cualquier cambio de plantilla derivado del cambio tecnológico será objeto de consulta y acuerdo, con garantías de recolocación, recualificación o indemnización adecuada.
— Si fuera posible, establecer un período en el que el sistema antiguo y el nuevo operen en paralelo, lo que permitiría verificar que el nuevo sistema no introduce nuevos sesgos o problemas antes de retirar el anterior completamente. Durante ese período, las decisiones del nuevo sistema pueden someterse a una supervisión más estricta, o incluso no aplicarse si hay discrepancias graves con el sistema anterior.

La fase de desmantelamiento supone un punto ciego en los análisis de IA laboral. Existe mucha literatura sobre cómo *entra* el algoritmo en la empresa, pero muy poca sobre cómo *sale* y qué garantías deben quedar activas tras su salida[13]. En otras palabras, la literatura analiza el "despido *por* algoritmo", pero raramente el "despido *del* algoritmo". Sin embargo, es un momento en el que la gobernanza algorítmica debe reforzarse para evitar que el fin de un sistema implique el fin de la protección de los trabajadores y las trabajadoras que vivieron bajo él. Solo así el ciclo de vida del algoritmo en la empresa se cierra de forma ordenada, justa y respetuosa con los derechos que hayan sido adquiridos.

7. Conclusiones

La propuesta de una nueva agenda de gobernanza algorítmica basada en la intervención de la RLT a lo largo de un ciclo multifásico demuestra que es posible articular un modelo coherente en el que la autonomía colectiva no cede ante la lógica tecnológica, sino que la subordina a los principios estructurales del Derecho del Trabajo: la protección de la parte débil, la igualdad, la participación y la salud laboral.

No se trata por tanto meramente de reaccionar ante sistemas ya implantados o de ejercer control pasivo sobre decisiones adoptadas, sino de participar activamente en la definición de qué puede hacer el algoritmo, con qué datos, bajo qué límites y mediante qué garantías. Esta reconceptualización de la negociación colectiva como herramienta de co-configuración social de la tecnología constituye una aportación fundamental para el Derecho del Trabajo contemporáneo y amplía un campo de intervención que permanecía poco explorado.

No obstante, la ruta hacia esa gobernanza requiere un cambio de paradigma más profundo: abandonar la fe en la transparencia como solución suficiente. Saber ya no es suficiente; hay que poder intervenir. Esta es la clave de la gobernanza algorítmica laboral propuesta.

Ese cambio de paradigma se hace especialmente urgente ante la paradoja crítica identificada en este análisis: la posibilidad del cumplimiento formal sin equidad real. O lo que es lo mismo, la evidencia de que es posible que una empresa pueda estar formalmente en conformidad con la regulación vigente y, simultáneamente, estar utilizando un sistema que discrimina sistemáticamente a colectivos vulnerables. Evidencia de la que se concluye, además, un elemento crítico de la gobernanza que

13. De hecho, el propio Reglamento 2024/1689 centra la mayoría de sus obligaciones (transparencia, supervisión humana, calidad de datos) en el ciclo de vida activo del sistema. Tan sólo el artículo 9 (Gestión de Riesgos) menciona el ciclo de vida completo. No hay, sin embargo, un artículo específico en la norma que regule detalladamente los derechos de las personas trabajadoras durante el desmantelamiento (más allá de la obligación general de conservar logs durante un tiempo, recogida en el artículo 12).

aquí se articula: la necesidad de una evaluación de impacto laboral específica, distinta e integrada junto a la evaluación normativa que cierre esta brecha.

La operatividad de esta gobernanza depende también de una transformación radical del papel de la representación de las personas trabajadoras, que debe convertirse en un actor técnico cualificado, no meramente consultivo. Esta transformación no es un lujo, sino una exigencia de eficacia. Sin ella, la RLT negocia a ciegas y los pactos convencionales quedan vacíos. Es necesario reconocer que la gobernanza algorítmica requiere inversión en capacidad técnica de los representantes y que esa inversión es un componente imprescindible de cualquier modelo serio de regulación.

Tampoco puede caer en el olvido que los sistemas se retiran o se sustituyen en algún momento y que la gobernanza debe estar presente por tanto en el ciclo completo de vida del algoritmo, desde el diseño hasta el desmantelamiento. En otras palabras, articular una gobernanza que cierre este ciclo de forma ordenada -garantizando preservación de evidencias, negociación de transición y protección de derechos adquiridos- no debería ser una innovación doctrinal que diferencia este trabajo.

Con todo, es imprescindible reconocer que la propuesta enfrenta desafíos significativos de implementación. Exige transformación sustancial de prácticas convencionales de negociación; requiere que se resuelvan genuinos problemas técnicos de interpretabilidad; implica costes económicos considerables; y presupone una sensibilidad especial de la RLT hacia colectivos vulnerables que no siempre existe. Estos límites no invalidan el modelo, pero subrayan que su implementación requiere un compromiso sistémico. Estos es, cambios en legislación laboral, inversión en capacidad técnica de representantes, reformas en procedimientos de negociación, y una mayor conciencia del riesgo que representa dejar la IA laboral completamente en manos de lógica empresarial no regulada.

Por todo lo anterior, el futuro del trabajo bajo inteligencia artificial no está predeterminado. No es inevitable que los algoritmos sustituyan completamente el criterio humano en decisiones laborales críticas, ni que las personas trabajadoras sean sujetos pasivos de sistemas opacos. Pero existe una alternativa viable al cambio tecnológico en el trabajo y es la que pasa por colocar a la negociación colectiva y a RLT en el centro de la gobernanza algorítmica. La nueva agenda aquí propuesta es un modelo que permite que la tecnología sirva a los derechos laborales, en lugar de que los derechos se adapten a la ella. Lo que únicamente será posible si existe voluntad de todas las partes de que así sea.

Bibliografía y documentación

Dastin, Jeffrey (2018). "Amazon scraps secret AI recruiting tool that showed bias against women". *Reuters*. Publicado el 11/10/2018 (en línea). Disponible en: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/ [consulta: 03/10/2025].

Fabregat Monfort, Gemma (2024). "Procesos de selección algorítmica y discriminación". *LABOS. Revista de Derecho del Trabajo y Protección Social*, Vol. 5, pp. 131–153.

Ginès i Fabrellas, Anna (2024). "Analítica de personas y discriminación algorítmica en procesos de selección y contratación". *LABOS. Revista de Derecho del Trabajo y Protección Social*, Vol. 5, pp. 99–130.

Prince, Anya & Schwarcz, Daniel (2020). *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*. *Iowa Law Review*, Vol. 105, pp. 1257–1318.

Documentos institucionales:

ISO/IEC 23053 – *Framework for AI systems using machine learning*. Estándar internacional utilizado como referencia para la estructuración del ciclo de vida de los sistemas de IA.

CEN-CENELEC JTC 21 – *AI Standards: Complete Detailed Overview*. Documento europeo de referencia para reguladores, industria y actores vinculados a la implementación del Reglamento Europeo de IA.

Ministerio de Trabajo y Economía Social (España) (2022) – *Guía práctica sobre información algorítmica en el ámbito laboral*.

Supervisor Europeo de Protección de Datos (EDPS) (2023) – *TechDispatch: Explainable Artificial Intelligence*. Publicado el 16/11/2023 (en línea). Disponible en: https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf [Consulta: 06/11/2025].

AEPD (2021/2024) – *Guía de Auditoría de Tratamientos que incluyan IA*. Publicado en enero de 2021 (en línea). Disponible en: https://www.aepd.es/guias/requisitos-auditorias-tratamientos-incluyan-ia.pdf**).** [Consulta: 18/11/2025].

Informes y literatura especializada:

Digital Future Society (2022) – *La discriminación algorítmica en España: límites y potencial del marco legal*. Publicado en septiembre de 2022 (en línea). Disponible en: https://digitalfuturesociety.com/app/uploads/2022/09/Discriminacion_algoritmica_Espana_marco_legal.pdf [consulta: 03/12/2025]

IBM (2025) – *La IA explicable*. Think Report. *Disponible en:* https://www.ibm.com/es-es/think/topics/explainable-ai [Consulta: 06/11/2025].