
This is the **published version** of the master thesis:

Salvador Mesa, Nil; Suquet Capdevila, Josep (tut.). *La prevenció de la discriminació algorítmica en els sistemes d'IA d'alt risc : una anàlisi dels requisits de l'AI Act*. (Universitat Autònoma de Barcelona), 2026 (Màster Universitari en Integració Europea)

This version is available at <https://ddd.uab.cat/record/331597>

under the terms of the  license.

UAB

Universitat Autònoma de Barcelona

MÀSTER UNIVERSITARI EN INTEGRACIÓ EUROPEA

Curs: 2025-2026

TREBALL DE FI DE MÀSTER

**La prevenció de la discriminació algorítmica en els sistemes
d'IA d'alt risc: una anàlisi dels requisits de l'AI Act**

Autor:

Nil Salvador Mesa

Professor:

Dr. Josep Suquet Capdevila

Facultat de Dret, 18 de juny de 2026

Resum:

La creixent utilització de sistemes d'intel·ligència artificial en processos de presa de decisions planteja riscos discriminatoris que poden tenir un impacte significatiu en els drets i oportunitats de les persones. El present treball examina la capacitat dels requisits aplicables als sistemes d'IA d'alt risc previstos en l'*AI Act* per prevenir aquests riscos i protegir el principi d'igualtat. L'anàlisi mostra que la discriminació algorítmica acostuma a manifestar-se en forma de discriminació indirecta i opaca, fet que justifica l'adopció, per part de l'*AI Act*, d'un conjunt de mecanismes preventius orientats a reduir aquests riscos abans que es materialitzin en resultats discriminatoris.

Paraules clau: intel·ligència artificial (IA), *Artificial Intelligence Act (AI Act)*, sistemes d'IA d'alt risc, discriminació algorítmica, transparència algorítmica.

Resumen:

La creciente utilización de sistemas de inteligencia artificial en procesos de toma de decisiones plantea riesgos discriminatorios que pueden tener un impacto significativo en los derechos y oportunidades de las personas. El presente trabajo examina la capacidad de los requisitos aplicables a los sistemas de IA de alto riesgo previstos en el AI Act para prevenir estos riesgos y proteger el principio de igualdad. El análisis muestra que la discriminación algorítmica tiende a manifestarse en forma de discriminación indirecta y opaca, lo que justifica la adopción, por parte del AI Act, de un conjunto de mecanismos preventivos orientados a reducir estos riesgos antes de que se materialicen en resultados discriminatorios.

Palabras clave: inteligencia artificial (IA), *Artificial Intelligence Act (AI Act)*, sistemas de IA de alto riesgo, discriminación algorítmica, transparencia algorítmica.

Abstract:

The increasing use of artificial intelligence systems in decision-making processes raises discriminatory risks that may significantly affect individuals' rights and opportunities. This study examines the capacity of the requirements applicable to high-risk AI systems under the AI Act to prevent such risks and protect the principle of equality. The analysis shows that algorithmic discrimination tends to manifest itself through indirect and opaque forms of discrimination, which justifies the adoption, by the AI Act, of a set of preventive mechanisms aimed at reducing these risks before they materialise into discriminatory outcomes.

Keywords: artificial intelligence; AI Act; high-risk AI systems; algorithmic discrimination; algorithmic transparency.

SUMARI

I. Introducció i metodologia.....	3
II. Marc teòric: Intel·ligència artificial, discriminació algorítmica i igualtat.....	5
1. Sistemes d'intel·ligència artificial i algoritmes.....	5
1.1. Intel·ligència artificial i aprenentatge automàtic.....	5
1.2. Opacitat algorítmica.....	6
2. El principi d'igualtat com a valor fonamental de la UE.....	7
2.1. La igualtat i la no-discriminació de l'article 2 del TUE.....	7
2.1.1. La consolidació jurisprudencial de l'article 2 del TUE.....	7
2.2. Discriminació algorítmica.....	9
III. Els sistemes d'IA d'alt risc i els mecanismes de prevenció de la discriminació en l'AI Act.....	10
1. Sistemes d'intel·ligència artificial d'alt risc.....	11
1.1. Concepte i classificació.....	11
1.2. Riscos discriminatoris associats als sistemes d'alt risc.....	12
1.2.1. Biaixos derivats de les dades.....	12
1.2.2. Inferència estadística i criteris aparentment neutres.....	13
1.2.3. Retroalimentació i amplificació de desigualtats.....	14
1.2.4. Opacitat algorítmica i dificultat de detecció.....	15
2. Els requisits dels sistemes d'IA d'alt risc com a mecanismes de prevenció de la discriminació.....	16
2.1. Governança de dades i prevenció de biaixos.....	16
2.2. Transparència, documentació i traçabilitat.....	18
2.3. Supervisió humana.....	20
Conclusions.....	22
Legislació.....	23
Jurisprudència.....	24
Bibliografia.....	24

I. Introducció i metodologia

En els darrers anys, la intel·ligència artificial (IA) ha acabat esdevenint una tecnologia central en el funcionament de les societats contemporànies. Actualment aquesta tecnologia ja no s'utilitza únicament en àmbits tècnics; els sistemes d'IA han anat adquirint una presència creixent en els processos de presa de decisions en diversos àmbits socialment sensibles, fet que evidencia que ja no és només una eina tècnica, sinó que s'ha convertit en un element amb una clara dimensió política i social.

A causa d'aquesta progressiva proliferació, l'exercici del poder de decisió ha viscut una profunda transformació. En molts casos, la valoració de perfils, la classificació d'individus o la predicció de comportaments es delega a sistemes algorítmics d'aprenentatge automàtic i un dels reptes principals associats a aquest fenomen és la dificultat que suposa comprendre com funcionen en la pràctica.

Aquesta dificultat s'explica, en gran mesura, per la manca de transparència que caracteritza aquests sistemes, la qual dificulta la capacitat de supervisar, qüestionar o impugnar les decisions automatitzades. Això no només representa un repte tècnic, sinó que també suposa grans desafiaments des del punt de vista de la governança i la legitimitat d'aquestes decisions delegades a la IA.

Per entendre el perquè de la gravetat de la manca de transparència en aquests processos de presa de decisions, cal tenir en compte que els sistemes d'aprenentatge automàtic es nodreixen de dades produïdes en contextos socials amb biaixos preexistents i, en conseqüència, la IA pot reproduir o intensificar aquestes desigualtats estructurals.

El problema de tot això resideix en l'aparença de neutralitat que poden adquirir els resultats algorítmics a causa de l'absència de mecanismes de detecció dels biaixos preexistents –o dels que es generen durant el procés intern dels sistemes–, fet que pot dificultar la detecció de les discriminacions en aquests resultats i contribuir a la seva normalització.

En aquest context, el 21 d'abril de 2021 la Comissió Europea –conscient dels grans avantatges de la IA, però també dels potencials riscos– va fer una proposta per establir un marc jurídic amb normes harmonitzades en matèria d'IA: el Reglament d'Intel·ligència Artificial (d'ara en endavant *AI Act*, reglament o reglament (UE) 2024/1689).

L'aprovació de la proposta al desembre de 2023 per part del Parlament Europeu i el Consell de la UE –i la conseqüent entrada en vigor a l'agost de 2024– respon a la voluntat de la UE, com a potència normativa, de posicionar-se com a referent normatiu a escala global en matèria d'IA. “La UE aspira a ser el líder mundial en IA segura. Desenvolupant un marc regulador sòlid basat en els drets humans i els valors fonamentals, la UE pot desenvolupar un ecosistema d'IA que beneficiï tothom.” (Comissió Europea, 2024; traducció pròpia).

L'arquitectura d'aquest reglament, com bé defineix la Comissió Europea (2021) es basa en una sèrie de components bàsics que, en conjunt, construeixen un enfocament normatiu europeu proporcionat i basat en el risc. Dins aquesta arquitectura, els sistemes classificats com “d'alt risc” ocupen una posició central, atès que són aquells que es considera que poden incidir de manera més directa sobre els drets i oportunitats de les persones.

És per això que aquest treball sorgeix de l'interès per examinar el paper que poden jugar els mecanismes creats per protegir els drets i oportunitats de les persones davant els riscos de discriminació associats als sistemes d'IA. En el marc de l'*AI Act*, que introdueix certs

requisits específics per als sistemes classificats com d'alt risc, resulta especialment rellevant analitzar fins a quin punt aquestes mesures poden contribuir a prevenir o limitar els impactes discriminatoris de les decisions algorítmiques.

A partir d'aquestes consideracions, per marcar la línia general de l'estudi, el present treball es planteja la següent pregunta d'investigació:

- *En quina mesura els requisits aplicables als sistemes d'IA d'alt risc previstos a l'AI Act contribueixen a afrontar els riscos de discriminació associats a aquests sistemes per protegir la igualtat com a valor fonamental de la UE?*

Per tal de mantenir un enfocament acotat a la pregunta d'investigació i fidel a les limitacions d'extensió del treball, l'estudi es centrarà específicament en el Reglament (UE) 2024/1689 com a principal instrument normatiu europeu en matèria de regulació de la IA.

Dins aquest marc normatiu, l'anàlisi es limitarà als requisits que estableix aquest reglament per als sistemes classificats com d'alt risc, ja que com s'ha explicat prèviament, el mateix *AI Act* els identifica com aquells que poden afectar més significativament els drets de les persones. En particular, el treball es centrarà exclusivament en els riscos de discriminació associats a aquests sistemes, en la mesura que poden incidir directament en el principi d'igualtat com a valor fonamental de la Unió Europea, recollit a l'article 2 del TUE.

L'objectiu principal del present treball és analitzar, des d'una perspectiva conceptual i normativa, la capacitat dels requisits aplicables als sistemes d'IA d'alt risc, previstos a l'*AI Act*, per prevenir i limitar els riscos de discriminació associats a aquests sistemes. En conseqüència –i tenint en compte les limitacions d'extensió–, no s'abordarà de manera exhaustiva altres instruments normatius relacionats amb la regulació de la IA, ni tampoc s'inclourà l'anàlisi empírica de casos concrets d'aplicació de decisions algorítmiques¹.

Metodològicament, per tal de realitzar el treball de manera coherent amb l'objecte d'estudi que s'acaba de delimitar, s'adoptarà un enfocament basat en l'anàlisi normativa del marc regulador europeu mitjançant la qual s'examinaran, en concret, les disposicions recollides a la secció 2 del Capítol III de l'*AI Act* relatives als requisits aplicables als sistemes d'IA d'alt risc. Entre aquests requisits, s'analitzaran especialment els vinculats a la governança de dades, la transparència, la traçabilitat i la supervisió humana.

Aquest marc d'anàlisi es complementarà amb la revisió bibliogràfica de textos acadèmics sobre transparència algorítmica, discriminació algorítmica i governança de la IA, per així poder situar l'anàlisi normativa dins el debat acadèmic existent sobre aquesta qüestió.

A partir d'aquest enfocament, el treball s'estructurarà en diversos apartats amb l'objectiu últim de donar resposta a la pregunta d'investigació que s'ha plantejat. Es començarà amb un primer capítol on es desenvoluparà el marc teòric necessari per la realització de l'estudi, en el que s'introduiran els principals conceptes relacionats amb la IA, així com les reflexions que s'han considerat més rellevants sobre el principi d'igualtat i la no-discriminació.

Es continuarà amb un segon capítol en el que s'analitzarà el marc normatiu europeu en matèria d'IA, centrant-se específicament en l'*AI Act* i, en particular, en els requisits aplicables als sistemes classificats com d'alt risc. Finalment, a les conclusions del treball, es valorarà la

¹ Cal destacar que, tot i no analitzar-se exhaustivament, si és necessari per a una millor comprensió de l'estudi, sí que s'utilitzaran tant disposicions d'altres instruments normatius com exemples reals de discriminació algorítmica, per donar consistència a l'argument central.

capacitat d'aquests requisits per prevenir els riscos discriminatoris associats als sistemes d'IA d'alt risc i les seves implicacions en la protecció del principi d'igualtat.

II. Marc teòric: Intel·ligència artificial, discriminació algorítmica i igualtat

1. Sistemes d'intel·ligència artificial i algoritmes

Per iniciar aquesta contextualització teòrica, es considera oportú començar aclarint què s'entén per intel·ligència artificial. Tenint en compte que el present treball es centra específicament en l'*AI Act*, és pertinent partir de la definició que es proposa en aquest mateix reglament. L'article 3 del *AI Act* defineix a un sistema d'IA de la següent manera:

un sistema basat en una màquina que està dissenyat per funcionar amb diferents nivells d'autonomia i que pot mostrar capacitat d'adaptació després del desplegament, i que, per objectius explícits o implícits, infereix de la informació d'entrada que rep la manera de generar resultats de sortida, com ara prediccions, continguts, recomanacions o decisions, que poden influir en entorns físics. (Reglament (UE) 2024/1689, article 3; traducció pròpia)

La definició de sistema d'IA que s'acaba de proporcionar ens permet saber que aquests sistemes es caracteritzen per la seva capacitat de processar informació i generar resultats que poden tenir un impacte en diversos entorns –ja siguin físics o virtuals– que, en el que respecta l'interès del present treball, en aquests entorns també s'hi inclou la presa de decisions. Tenint en compte això, cal destacar que el mecanisme central d'aquest funcionament tan característic dels sistemes d'IA és l'aplicació d'algoritmes, que permeten a aquests sistemes analitzar les dades que reben, identificar patrons i produir resultats.

La RAE defineix un algoritme com “un conjunt ordenat i finit d'operacions que permet trobar la solució d'un problema”. En l'àmbit dels sistemes d'IA, els algoritmes estan integrats dins de sistemes algorítmics, una estructura més complexa que combina l'ús d'algoritmes amb altres components tècnics i organitzatius per processar les dades i generar resultats. En aquest sentit, Bandy (2021; traducció pròpia) defineix el sistema algorítmic com qualsevol sistema sociotècnic influenciat per almenys un algoritme. Això inclou sistemes que poden dependre del judici humà i/o altres components no algorítmics, sempre que incloguin almenys un algoritme.

En les darreres dècades, els sistemes algorítmics han experimentat una evolució significativa en relació a la seva capacitat de millorar progressivament el rendiment a partir de la informació processada, gràcies al desenvolupament de noves tècniques.

1.1. Intel·ligència artificial i aprenentatge automàtic

Entre aquestes noves tècniques que s'acaben de mencionar, destaca especialment el que es coneix com aprenentatge automàtic (*machine learning*), el qual permet als algoritmes identificar patrons en grans conjunts de dades i millorar el seu rendiment a partir de l'experiència acumulada.

Es considera rellevant parlar de l'aprenentatge automàtic perquè, actualment, la gran majoria de sistemes d'IA estan constituïts per algoritmes que funcionen mitjançant aquesta tècnica. Sobre aquesta base, destaquen dues qüestions centrals a tenir en compte sobre el *machine learning*: de què aprenen i com aprenen?

En relació a la primera qüestió, destaca la definició que fa Rouhiainen (2018) de la IA: “podríem dir que la IA és la capacitat de les màquines per utilitzar algoritmes, aprendre de les dades i utilitzar allò que s’ha après en la presa de decisions tal com ho faria un ésser humà” (traducció pròpia). És a dir, el primer punt a tenir en compte és que l’aprenentatge automàtic fa que els algoritmes –i conseqüentment els sistemes d’IA– aprenguin de les dades que van rebent. Això és important ja que els hi permet millorar el seu rendiment de manera autònoma, cosa que els hi permet no haver d’anar sent programats per a cada tasca concreta.

A continuació, per respondre a la segona qüestió, es farà una breu descripció de com funcionen els tres models diferents d’aprenentatge automàtic existents: el supervisat, el no supervisat i per reforç.

L’aprenentatge supervisat és un model d’aprenentatge automàtic que utilitza dades que ja tenen la resposta correcta associada. Per entrenar aquest tipus de model s’utilitzen dues categories de dades, en primer lloc, les dades d’entrenament, amb les que l’algoritme aprèn la relació entre les dades d’entrada i el resultat esperat. Després, s’utilitzen les dades de prova, amb les que s’examina si el model és capaç de fer prediccions correctes amb informació nova, és a dir, dades que no ha vist durant l’entrenament.

En el cas del model d’aprenentatge no supervisat, el sistema “s’entrena en funció de les característiques de les dades sense etiquetar i després es pren una decisió sobre les dades de prova disponibles” (Pandey, Niwaria i Chourasia; 2019; traducció pròpia). En altres paraules, en lloc d’aprendre a partir de respostes conegudes, el model analitza les característiques de les dades rebudes, observant les similituds i les diferències, per identificar patrons, relacions o estructures ocultes dins del conjunt de dades i, així, poder agrupar-les o organitzar-les d’una determinada manera.

Finalment, “el model d’aprenentatge per reforç es basa exclusivament en dues funcions: la cerca per assaig i error, i el resultat retardat.” (Pandey, Niwaria i Chourasia; 2019; traducció pròpia). És a dir, l’algoritme pren decisions de manera progressiva i ajusta el seu comportament, en funció del *feedback* que se li dona, per apropar-se a l’objectiu. Per tant, l’aprenentatge per reforç es diferencia dels models anteriors en la mesura que l’algoritme aprèn directament a partir de la interacció amb l’entorn.

1.2. Opacitat algorítmica

Com es pot observar amb el que s’acaba d’exposar en l’apartat anterior, els models d’aprenentatge automàtic permeten als sistemes d’IA desenvolupar comportaments cada cop més complexos a partir de les dades que recullen i de la seva interacció amb l’entorn.

La creixent complexitat del comportament dels sistemes d’IA, a la vegada, comporta una major dificultat per comprendre com es generen els resultats a la pràctica, és a dir, per identificar quins processos interns ha seguit el sistema per produir l’*output*.

Aquest obstacle per comprendre els processos interns dels sistemes d’IA es deu al que es coneix com a opacitat algorítmica. És important destacar que l’opacitat algorítmica no s’ha d’entendre únicament com la dificultat tècnica d’aquests sistemes, ja que és un fenomen complex format per diversos factors estretament connectats.

Aquests factors són la manca de transparència, de documentació i de traçabilitat, els quals s’hi entrarà profundament en detall en l’apartat 2.1 del capítol III. En aquest context, la transparència no s’ha d’entendre en sentit estricte d’accés a la informació sinó com un

mecanisme orientat a combatre l'opacitat algorítmica en totes les seves dimensions mencionades abans, és a dir, no només pel que fa la manca de transparència, sinó també en relació a la manca de documentació i traçabilitat.

2. El principi d'igualtat com a valor fonamental de la UE

En relació al tema central del present estudi, l'opacitat algorítmica esdevé un dels principals problemes a analitzar, ja que pot dificultar la detecció de possibles biaixos presents en les dades d'entrada o en el propi funcionament dels sistemes algorítmics, i en conseqüència, generar resultats discriminatoris que impactin de manera perjudicial contra el principi d'igualtat.

2.1. La igualtat i la no-discriminació de l'article 2 del TUE

El principi d'igualtat, segons l'article 2 del Tractat de la Unió Europea², constitueix un dels valors fonamentals sobre els quals es sustenta la Unió. El mateix article, un cop establerts els valors fonamentals, concreta algunes manifestacions d'aquests valors en forma de principis, els quals –segons l'article– caracteritzen les societats dels Estats membres. Entre les característiques recollides, el valor d'igualtat es materialitza en la no-discriminació i la igualtat entre dones i homes.

Com s'ha anat esmentant, en el context dels sistemes d'IA, el principi mencionat adquireix una rellevància especial, en la mesura que aquests sistemes intervenen cada cop més en els processos de presa de decisions, les quals poden acabar esdevenint discriminatòries sobre determinats col·lectius vulnerables.

La possible existència de biaixos en aquests sistemes planteja la necessitat d'analitzar fins a quin punt aquests poden incidir negativament en el respecte al principi d'igualtat. Per tant, l'anàlisi del principi d'igualtat en aquest treball no es planteja en termes abstractes, sinó en relació amb els riscos específics de discriminació derivats de l'ús dels sistemes d'IA.

2.1.1. La consolidació jurisprudencial de l'article 2 del TUE

Abans de continuar amb l'anàlisi central de l'estudi, és important destacar que l'article 2 del TUE ha adquirit una rellevància jurídica especialment significativa en els últims anys, fet que es troba vinculat a la recent evolució de la jurisprudència del Tribunal de Justícia de la Unió Europea (TJUE).

En un primer moment, els valors recollits en aquest precepte es van configurar com a principis de caràcter aspiracional i programàtic, orientats sobretot a reforçar la identitat del projecte europeu i projectar a l'exterior la imatge de la UE com una comunitat fonamentada en valors comuns i drets fonamentals³. Així doncs, jurídicament aquest article actuava més aviat com una disposició amb una funció de suport i per interpretar altres normes, i no obligava per si mateix.

No obstant això, davant les creixents tensions en relació a la protecció dels drets fonamentals i de l'Estat de Dret dins de la UE, aquesta concepció sobre l'article 2 del TUE ha canviat

² Article 2 del TUE:

La Unió es fonamenta en els valors de respecte de la dignitat humana, llibertat, democràcia, igualtat, Estat de Dret i respecte dels drets humans, inclosos els drets de les persones pertanyents a minories. Aquests valors són comuns als Estats membres en una societat caracteritzada pel pluralisme, la no-discriminació, la tolerància, la justícia, la solidaritat i la igualtat entre dones i homes.

³ És a dir, de reforçar la seva imatge davant la resta d'actors internacionals com a potència normativa.

notablement en els darrers anys. El TJUE ha anat reconeixent una major força normativa als valors recollits en aquest precepte, fet que ha comportat que hagin deixat de ser considerats únicament com a principis orientadors, per també passar a tenir una eficàcia jurídica real, autònoma i vinculant.

En aquest context, resulta especialment rellevant la sentència C-769-22 del TJUE en relació a l'assumpte Comissió c. Hongria (vegeu l'apartat de jurisprudència). La Comissió Europea va interposar un recurs per incompliment contra Hongria com a resposta davant l'adopció d'aquest país de la *A pedofil bűnelkövetőkkel szembeni szigorúbb fellépésről, valamint a gyermekek védelme érdekében egyes törvények módosításáról szóló 2021. évi LXXIX. törvény* (Llei LXXIX de 2021, relativa a l'enduriment de les mesures contra els delinqüents pedòfils i a la modificació de determinades lleis per tal de garantir la protecció dels infants) en l'ordenament jurídic hongarès, l'any 2021.

Formalment, aquesta reforma legislativa estava orientada a protegir els menors d'edat, però diverses de les disposicions introduïdes restringien i/o prohibien l'accés dels menors a continguts que representessin o promoguessin l'homosexualitat o la identitat de gènere diferent a la del sexe assignat al néixer. Entre altres coses, la normativa limitava la difusió d'aquests tipus de continguts en l'àmbit educatiu, audiovisual i publicitari⁴.

Alhora, conjuntament a aquestes restriccions, la mateixa reforma legislativa introduïa l'enfortiment d'algunes mesures contra la pedofília i determinats delictes sexuals contra menors, cosa que, segons la Comissió –i posteriorment el TJUE– generava una associació implícita entre homosexualitat, identitat de gènere i conductes sexuals il·lícites. En relació amb l'adopció d'aquesta llei, la Comissió va formular sis imputacions basades en infraccions del dret europeu –tant de l'originari com del derivat–⁵.

A part de l'incompliment de diverses directives europees, la Comissió va posar especialment l'accent en l'incompliment de l'article 21⁶ de la Carta dels Drets Fonamentals de la Unió Europea, el qual prohibeix expressament qualsevol discriminació basada, entre altres raons, en l'orientació sexual. També destaca la imputació basada en la infracció dels principis recollits en l'article 2 del TUE, –especialment els principis d'igualtat, dignitat humana i respecte pels drets humans– ja que la Comissió al·legava que aquesta reforma legislativa projectava una estigmatització institucionalitzada de les persones LGTBI, cosa que representava un acte incompatible amb els valors sobre els quals es fonamenta la UE.

En la seva sentència, el TJUE va estimar substancialment el recurs interposat per la Comissió i va declarar que Hongria havia incomplert el dret de la Unió. El tribunal va considerar que les restriccions que imposava la normativa hongaresa discriminaven segons l'orientació sexual i, per tant, vulneraven tant la Carta dels Drets Fonamentals de la UE com obligacions provinents del dret derivat de la Unió.

⁴ Un exemple d'aquestes limitacions és que la normativa impedia que determinats programes de televisió dirigits a menors abordessin qüestions relatives a la diversitat sexual o de gènere.

⁵ “En suport del seu recurs, la Comissió formula sis imputacions, basades, la primera, en infraccions dels articles 2, 3, apartat 1, 6 bis, apartat 1, i 9, apartat 1, lletra c), incís ii), de la Directiva 2010/13; la segona, en una infracció de l'article 3, apartat 2, de la Directiva 2000/31; la tercera, en infraccions dels articles 16 i 19 de la Directiva 2006/123 i de l'article 56 TFUE; la quarta, en una infracció de l'article 10 del RGPD i de l'article 8, apartat 2, de la Carta; la cinquena, en infraccions dels articles 1, 7, 11 i 21 de la Carta, i, la sisena, en una infracció de l'article 2 TUE”. (Tribunal de Justícia de la UE, 2022; traducció pròpia)

⁶ Article 21, Carta de Drets Fonamentals de la Unió Europea (traducció pròpia):

Es prohibeix tota discriminació i, en particular, l'exercida per raó de sexe, raça, color, orígens ètnics o socials, característiques genètiques, llengua, religió o conviccions, opinions polítiques o de qualsevol altre tipus, pertinença a una minoria nacional, patrimoni, naixement, discapacitat, edat o orientació sexual.

A més, aquesta resolució adquireix una major transcendència perquè és quan, per primer cop, es declara de manera tan directa i autònoma que un Estat Membre incomplia l'article 2, ja que el tribunal, en la mateixa sentència, va afirmar que la vulneració dels valors fonamentals de la Unió pot constituir, per si mateixa, una infracció del dret de la UE.

2.2. Discriminació algorítmica

El principi de no-discriminació no representa un principi separat al d'igualtat, sinó que s'integra dins aquest, sent el primer una manifestació qualificada del segon, en la mesura que la no-discriminació projecta el principi d'igualtat de manera reforçada sobre determinats factors especialment sensibles⁷. És a dir, tota discriminació implica una vulneració del principi d'igualtat, però no tota vulneració de la igualtat és discriminació.

Aquesta precisió conceptual resulta necessària, ja que per raons d'extensió, el present treball es centrarà en aquells supòsits en què la vulneració del principi d'igualtat s'hagi donat per causes discriminatòries derivades dels resultats generats pels sistemes algorítmics.

La discriminació algorítmica s'entén des de la perspectiva de que aquests sistemes poden utilitzar diferents fonts de dades per dur a terme les seves decisions, com ara dades històriques, patrons de comportament o dades d'enquestes, entre d'altres. Tot i això, aquestes dades poden contenir biaixos que es tradueixen en discriminació, cosa que es tradueix en l'exclusió de certs grups de persones, com minories ètniques, immigrants, dones o persones amb discapacitat, entre d'altres. (Danesi, 2022; citat per Iturmendi Rubia, 2023; traducció pròpia).

És important destacar que jurídicament es distingeixen dos tipus de discriminació: la discriminació directa i la discriminació indirecta. La *Directiva 2006/54/CE del Parlament Europeu i del Consell, de 5 de juliol del 2006, relativa a l'aplicació del principi d'igualtat d'oportunitats i igualtat de tracte entre homes i dones en assumptes treball i ocupació* defineix la discriminació directa com “la situació en què una persona sigui, hagi estat o pugués ser tractada per raó de sexe de manera menys favorable que una altra en situació comparable” (traducció pròpia).

Per altra banda, la mateixa directiva defineix la discriminació indirecta com:

la situació en què una disposició, criteri o pràctica aparentment neutres situen persones d'un sexe determinat en desavantatge particular pel que fa a persones de l'altre sexe, llevat que aquesta disposició, criteri o pràctica es pugui justificar objectivament amb una finalitat legítima i que els mitjans per assolir aquesta finalitat siguin adequats i necessaris⁸. (traducció pròpia)

Aquesta distinció és essencial en el context dels sistemes d'IA, ja que permet entendre una mica millor la complexitat del problema de la discriminació algorítmica. Basant-nos en les definicions donades per la pròpia Unió Europea, podem concloure que la majoria de situacions en què es produeix la discriminació algorítmica, es donen en forma de discriminació indirecta. Això es deu, com s'ha anat exposant al llarg del treball, a l'opacitat que caracteritza els sistemes algorítmics –especialment els d'aprenentatge automàtic–, mentre que la incorporació explícita de criteris discriminatoris resultaria més fàcilment detectable.

⁷ Aquests factors especialment sensibles recollits generalment per igual en la majoria dels ordenaments jurídics són: naixement, raça, sexe, religió, opinió o qualsevol altra condició o circumstància personal o social.

⁸ En aquest cas s'està parlant d'un factor especialment sensible en concret, però la definició –tan de la discriminació directa com de la indirecta– és extrapolable a tots els altres factors especialment sensibles detallats a la nota de peu de pàgina anterior.

És a dir, atès que el dret de la UE –i el dels Estats membres– prohibeix de manera expressa la discriminació per raons de naixement, sexe, raça, religió, entre altres (vegeu nota de peu de pàgina número 7), la inclusió explícita d'aquests criteris en el disseny d'un algoritme seria, en principi, fàcilment detectable i, conseqüentment, sancionable. En canvi, la discriminació indirecta pot derivar de l'ús de variables aparentment neutres o de biaixos preexistents de les dades treballades, fet que la fa més difícil de detectar abans que es generin els resultats, ja que tot això succeeix en els processos interns mencionats.

Aquesta dificultat per detectar els efectes discriminatoris abans que es materialitzin amb els *outputs*, evidencia l'existència de riscos associats a l'ús de sistemes d'IA. Tal i com s'ha anat exposant al llarg d'aquest marc teòric, els riscos discriminatoris associats a aquests sistemes troben l'origen en factors com la recepció i ús de dades esbiaixades, el propi funcionament del sistema i l'opacitat dels sistemes.

Tanmateix, la potencialitat d'aquests riscos no es manifesta amb la mateixa intensitat en tots els sistemes d'intel·ligència artificial. És per això, que quan la Unió Europea va decidir regular expressament la IA, ho va fer des d'un enfocament basat en el risc, diferenciant així els sistemes d'IA segons el nivell de risc d'impacte sobre els drets i llibertats de les persones. En aquest context, el següent capítol es centrarà en l'anàlisi dels sistemes d'IA d'alt risc i en els riscos discriminatoris que s'hi associen, en el marc de l'*AI Act*.

III. Els sistemes d'IA d'alt risc i els mecanismes de prevenció de la discriminació en l'*AI Act*

En matèria d'intel·ligència artificial, abans de la redacció de l'*AI Act*, no existia una normativa específica i centralitzada que regulés de manera expressa la IA. Per contra, el funcionament i impacte dels sistemes d'IA es trobaven regulats de manera fragmentada en diferents instruments jurídics europeus⁹.

Veient el creixement exponencial de l'ús dels sistemes d'IA, l'any 2024 la Unió Europea va aprovar l'*AI Act* amb l'objectiu de:

millorar el funcionament del mercat interior mitjançant l'establiment d'un marc jurídic uniforme, en particular per al desenvolupament, la introducció al mercat, la posada en servei i la utilització de sistemes d'intel·ligència artificial (d'ara endavant, «sistemes d'IA») a la Unió, de conformitat amb els valors de la Unió, per tal de promoure l'adopció d'una intel·ligència artificial (IA) centrada en l'ésser humà i fiable, garantint alhora un elevat nivell de protecció de la salut, la seguretat i els drets fonamentals consagrats a la Carta dels Drets Fonamentals de la Unió Europea (d'ara endavant, «Carta»), inclosos la democràcia, l'Estat de Dret i la protecció del medi ambient, protegir davant dels efectes perjudicials dels sistemes d'IA a la Unió, així com donar suport a la innovació. (Reglament (UE) 2024/1689; traducció pròpia)

Per establir aquesta regulació harmonitzada de normes sobre la IA es va plantejar un enfocament basat, segons el mateix reglament, en la intensitat i abast dels riscos que poden generar aquests sistemes sobre els drets fonamentals, la seguretat i la salut de les persones. En relació amb aquest enfocament, la transparència –entesa de la manera en la que s'ha definit en l'apartat [1.2](#) del Capítol II– es configura com un element central d'aquest reglament.

⁹ Alguns dels instruments jurídics que regulaven aspectes relatius a la IA o que es podien aplicar a la IA eren: el Reglament General de Protecció de Dades (RGPD); la Directiva 2006/54/CE del Parlament Europeu i del Consell que prohibeix la discriminació, ja sigui directa o indirecta; la Carta dels Drets Fonamentals de la UE amb els articles 20 i 21 (vegeu peus de pàgina 4 i 5); el *Digital Services Act* (DSA) que vetlla perquè l'espai digital respecti els drets fonamentals dels ciutadans i consumidors; el *Digital Market Act* (DMA) orientat a limitar el poder dels *gatekeepers*; la *Data Governance Act* (DGA) centrada en la governança i la reutilització de dades.

1. Sistemes d'intel·ligència artificial d'alt risc

1.1. Concepte i classificació

Tal com s'acaba d'esmentar, l'*AI Act* adopta un enfocament normatiu basat en el risc i, a partir d'aquest plantejament, el reglament estableix una classificació dels sistemes d'IA segons el seu potencial nivell de risc.

Aquest sistema classifica els sistemes d'IA en quatre nivells: inacceptable (que comporta prohibició absoluta), alt risc (subjecte a avaluació ex ante, amb supervisió humana), limitat/transparent (amb exigències bàsiques d'informació a l'usuari) i mínim (sense regulació específica). (Banchio, 2025)

Els sistemes d'IA de risc inacceptable estan previstos a l'article 5 del reglament, el qual distingeix diversos tipus d'aplicacions de la IA que són incompatibles amb els valors de la UE i, per tant, queden estrictament prohibides¹⁰.

En segon lloc, trobem els sistemes d'IA d'alt risc, que són la categoria central del reglament, es troben regulats principalment a l'article 6. A diferència dels sistemes prohibits, els sistemes d'alt risc no queden exclosos del mercat, però sí que queden sotmesos a un règim jurídic especialment estricte, ja que poden afectar significativament els drets fonamentals, a les oportunitats i a les condicions de vida de les persones.

L'article 6 preveu dues vies perquè un sistema d'IA sigui considerat d'alt risc. El primer apartat de l'article fa referència als sistemes d'IA que formen part de productes regulats per la legislació sectorial harmonitzada de la UE recollida a l'Annex I, sobretot en àmbits vinculats a la seguretat dels productes. Pel que fa a la segona via, el segon apartat de l'article 6 considera d'alt risc determinats sistemes d'IA utilitzats en àmbits especialment sensibles previstos en l'Annex III.

No obstant l'existència d'aquestes dues categoritzacions de sistemes d'IA d'alt risc, l'anàlisi es centrarà principalment en els sistemes previstos a l'Annex III del reglament, ja que són els que presenten una major incidència potencial sobre els drets fonamentals i, en el que respecta el present treball, sobre el principi d'igualtat i la prohibició de la discriminació.

Aquesta decisió metodològica es deu al fet que, a diferència dels sistemes d'IA contemplats a l'Annex I –vinculats principalment a riscos relatius a la seguretat tècnica dels productes–, els sistemes d'IA inclosos a l'Annex III¹¹ operen en àmbits especialment sensibles en què els

¹⁰Aquestes aplicacions de la IA prohibides pel reglament són: manipulació subliminal, explotació de les vulnerabilitats de les persones perquè aquestes adoptin un comportament nociu, categorització de persones basant-se en característiques sensibles, puntuació social d'ús general, avaluació del risc de les persones de cometre un delictes futur basant-se en trets i característiques personals, inferir l'estat emocional d'una persona, identificació biomètrica remota a temps real.

¹¹ L'Annex III del *AI Act* preveu que, d'acord amb l'article 6.2, es consideren d'alt risc els sistemes d'IA que operin en els següents vuit àmbits descrits per Trail ML (2026):

1. Sistemes biomètrics i basats en la biometria (com ara la identificació biomètrica remota, la categorització de persones i els sistemes de reconeixement d'emocions).
2. La gestió i l'explotació d'infraestructures crítiques (com ara el trànsit rodat, el subministrament d'energia o les infraestructures digitals).
3. Educació i formació professional (com ara l'avaluació d'estudiants en centres educatius).
4. La gestió de l'ocupació i dels treballadors (com ara la contractació, l'avaluació del rendiment o l'assignació de tasques).
5. Accés a serveis i beneficis privats i públics essencials (com ara la qualificació creditícia, les avaluacions de riscos en les assegurances mèdiques i l'enviament de serveis d'emergència).
6. L'aplicació de la llei (com ara l'avaluació de la fiabilitat de les proves o l'anàlisi de delictes).

sistemes algorítmics poden influir de manera directa en la presa de decisions, que afecten les oportunitats, drets i condicions de vida de les persones.

Finalment, els sistemes d'IA de risc limitat previstos en el reglament estan sotmesos principalment a obligacions de transparència i, pel que fa als sistemes de risc mínim no hi ha previstes obligacions específiques més enllà del compliment general del dret de la Unió Europea.

1.2. Riscos discriminatoris associats als sistemes d'alt risc

Com s'ha esmentat en l'apartat [2.3](#) del capítol II, la discriminació algorítmica principalment es manifesta mitjançant formes de discriminació indirecta, ja que els sistemes d'IA no acostumen a incorporar criteris discriminatoris explícits que afectin els drets i les oportunitats de les persones, sinó que generen resultats discriminatoris a partir de dades esbiaixades, inferències estadístiques o criteris aparentment neutres.

Això adquireix una rellevància especial en relació amb els sistemes d'IA que operen en els àmbits previstos en l'Annex III. Tenint en compte el que s'acaba d'esmentar, el present apartat analitzarà els principals riscos discriminatoris associats als sistemes d'IA i per què pot ser tan problemàtic si es produeixen en els àmbits recollits en l'Annex III.

1.2.1. Biaixos derivats de les dades

Un dels principals riscos discriminatoris associats als sistemes d'IA d'alt risc deriva de les dades d'entrada amb les que s'alimenten aquests sistemes. Com molt bé diu Arochena (2024; traducció pròpia), la IA, com la intel·ligència humana, utilitza informació d'entrada per arribar a decisions, i la mateixa pot estar esbiaixada.

Aquest tipus de risc, tot i que es pot donar en tots els tipus de sistemes d'IA d'aprenentatge automàtic, resulta especialment freqüent en els sistemes d'aprenentatge supervisat, ja que aquests són entrenats a partir de dades etiquetades que incorporen resultats o decisions prèvies. En conseqüència, quan les dades d'entrenament estan prèviament esbiaixades¹², el sistema pot acabar aprenent i reproduint de manera automàtica en les seves prediccions o decisions, aquestes desigualtats històriques o patrons discriminatoris preexistents en les dades.

Aquest problema, com s'ha esmentat al principi de l'apartat, adquireix una rellevància especial en els sistemes d'aprenentatge automàtic que operen en els àmbits recollits en l'Annex III, ja que les decisions que es prenen en sectors com l'ocupació, l'educació o la provisió de serveis essencials, poden condicionar directament les oportunitats de les persones. Sobretot perquè, si s'utilitzen dades esbiaixades en les decisions preses per sistemes d'IA en aquests àmbits, pot provocar que es reproduixin desigualtats estructurals vinculades, entre altres factors, al gènere, l'origen ètnic o la situació socioeconòmica.

Un dels casos més famosos que exemplifica clarament el fenomen de reproducció de biaixos discriminatoris a partir de dades històriques, és el cas del sistema experimental de

7. La gestió de la migració, l'asil i el control de fronteres (com ara l'avaluació del risc per a la seguretat d'una persona o l'examen de les sol·licituds d'asil, visat o permisos de residència).

8. Administració de justícia i processos democràtics (com ara ajudar en la interpretació i la investigació de fets, lleis i l'aplicació de la llei o influir en les eleccions).

¹² “Dades discriminatòries, desactualitzades, obtingudes o tractades sense garantia de compliment de les lleis de protecció de dades” (Arochena, 2024).

contractació que va desenvolupar Amazon l'any 2014. Segons una investigació feta per Jeffrey Dastin, publicada a l'agència internacional de notícies Reuters l'any 2018, aquest sistema havia estat entrenat mitjançant currículums presentats a l'empresa durant un període de deu anys, amb la intenció d'analitzar i classificar els currículums nous que arribessin.

“Però el 2015, l'empresa es va adonar que el seu nou sistema no qualificava els candidats per a llocs de treball de desenvolupador de programari i altres llocs tècnics de manera neutra pel que fa al gènere” (Dastin, 2018; traducció pròpia), fet que es deu a que el sector tecnològic ha estat –i segueix estant– tradicionalment masculinitzat.

Com molt bé argumenta de Zárate Alcarazo (2023), el sistema havia estat entrenat amb els currículums de les persones que havien participat en processos de selecció similar anteriorment i que en molts casos havien estat seleccionades. És a dir, tenint en compte que es tracta d'un sector masculinitzat, les dades amb les que es va entrenar el sistema de contractació provenien majoritàriament d'homes.

La investigació publicada per Reuters revelava que, en conseqüència del que s'acaba d'esmentar, el sistema de contractació d'Amazon va aprendre per si mateix que els candidats de sexe masculí eren preferibles. És més, no només afavoria als candidats masculins sinó que penalitzava a les candidatures femenines, especialment, els currículums que incloïen paraules relacionades amb el sexe femení.

La investigació de Dastin (2018), publicada a Reuters, explica que des d'Amazon es va intentar modificar el sistema perquè fos neutre en aquest àmbit, però com era un model d'aprenentatge automàtic, no garantia que desenvolupés altres mecanismes discriminatoris de classificació. Finalment es va acabar dissolent l'equip responsable del projecte.

És a dir, el que fa rellevant aquest cas –i que el configura com a exemple perfecte per explicar el risc que s'està tractant en aquest subapartat– és que, tot i que el sistema no incorporava explícitament el gènere com a criteri de decisió, les dades històriques utilitzades durant el procés d'entrenament arrossegaven desigualtats preexistents que el sistema va acabar reproduint de manera automàtica.

1.2.2. Inferència estadística i criteris aparentment neutres

Un altre dels principals riscos de discriminació dels sistemes d'alt risc deriva de l'ús d'inferències estadístiques. A diferència dels biaixos derivats directament de les dades històriques, aquest risc està vinculat a l'establiment, per part dels sistemes algorítmics, de correlacions entre variables concretes i els resultats que han de predir o classificar.

Encara que no s'utilitzin explícitament categories protegides com el sexe, la raça o la religió, el sistema pot acabar fent servir variables *proxies*¹³ (variables indirectes) estretament vinculades amb aquestes categories.

Com en l'anterior cas, això resulta especialment rellevant en relació amb els sistemes previstos a l'Annex III de l'*AI Act* ja que, com molt bé argumenta Arochena (2024; traducció pròpia), “l'algoritme utilitza una variable propera, en principi aliena a una causa de discriminació, per inferir una resposta discriminatòria”, en el cas dels sistemes que s'estan analitzant, en àmbits que poden afectar drets i oportunitats de les persones.

¹³Les variables proxies s'entenen com aquells factors aparentment neutres, com poden ser el codi postal o determinats patrons de comportament, que li poden permetre al sistema algorítmic inferir indirectament informació relativa al gènere, a l'origen ètnic o a la situació socioeconòmica de les persones.

Un dels exemples més coneguts amb el que es pot veure la materialització d'aquest risc de discriminació és el sistema algorítmic COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*), utilitzat en el sistema judicial dels Estats Units d'Amèrica, més concretament, el cas *State v. Loomis* de la Cort Suprema de Wisconsin (2016).

Aquest sistema s'utilitzava per ajudar als jutges i altres autoritats a prendre decisions sobre les persones que estaven sotmeses a processos judicials en relació a les condemnes, la llibertat provisional, la llibertat condicional, i la supervisió posterior.

El 2016 una investigació va posar de manifest que la imparcialitat de les decisions judicials, recolzats per COMPAS, estava compromesa, entre altres, pel biaix racial. L'estudi assenyalava que les persones negres qualificades pel programa d'alt risc tenien gairebé el doble de possibilitats de no reincidir que els blancs amb la mateixa qualificació; per contra, les persones blanques etiquetades de baix risc tenien més probabilitats de reincidència que les negres igualment classificades. (Sánchez Martínez, 2025; traducció pròpia)

La investigació es va centrar en el cas *Loomis*, en el que un condemnat a sis anys de presó va veure condicionada la determinació de la seva pena a causa d'una avaluació sobre el seu risc de reincidència, basada en un informe generat pel sistema COMPAS. Tot i el debat que va generar i el risc existent que clarament comportava el sistema –especialment de discriminació racial–, el Tribunal Suprem de Wisconsin va acabar acceptant la constitucionalitat de l'informe, conclouent que no s'havien vulnerat ni les garanties del procés ni el principi d'igualtat.

És a dir, tot i que l'algoritme no utilitzava formalment la raça com a variable explícita, mitjançant correlacions estadístiques i variables indirectes, els resultats que acabava generant eren discriminatoris per determinats col·lectius racialitzats. Per tant, aquest segon tipus de risc evidencia que la neutralitat formal dels sistemes algorítmics no garanteix necessàriament la neutralitat material dels resultats que produeixen.

1.2.3. Retroalimentació i amplificació de desigualtats

Un altre dels riscos de discriminació associats als sistemes d'IA d'alt risc és la capacitat que tenen aquests sistemes per retroalimentar-se i amplificar les desigualtats preexistents. Aquest risc està estretament vinculat amb els altres dos riscos de discriminació que s'acaben d'analitzar ja que, tant els biaixos preexistents que poden contenir les dades com les inferències estadístiques esbiaixades del sistema, poden acabar incorporant-se de nou en el funcionament del sistema.

És a dir, això es produeix quan els resultats discriminatoris generats pel sistema passen a formar part de les noves dades que posteriorment utilitzarà el mateix sistema per continuar entrenant-se, ajustant-se o perfeccionant-se. Aquest fenomen pot ser molt perjudicial ja que pot comportar que es reforcin els patrons discriminatoris previs de manera progressiva i generar dinàmiques d'exclusió acumulativa sobre determinats col·lectius vulnerables en àmbits tan sensibles com són els que estan recollits en l'Annex III de l'*AI Act*.

Per mostrar la gravetat del risc que suposa aquest fenomen si es manifesta en un àmbit dels que es troben recollits en l'Annex III, resulta especialment rellevant explicar el cas del sistema SyRI (*System Risk Indication*). Aquest sistema algorítmic era utilitzat per les autoritats públiques dels Països Baixos, per detectar possibles fraus en matèria de prestacions

i ajudes socials a partir de l'anàlisi automatitzada de grans quantitats de dades personals i socioeconòmiques.

Això resultava especialment problemàtic perquè les classificacions generades per l'algoritme podien condicionar les futures dades utilitzades pel sistema i, per tant, acabar reforçant progressivament l'associació entre determinats perfils socioeconòmics i un major risc de frau, reproduint i intensificant automàticament patrons discriminatoris previs sobre col·lectius socialment vulnerables.

Finalment, la Cort de Districte de la Haya va concloure en la seva sentència [C/09/550982](#), que el sistema SyRI era contrari a l'article 8.2 del Conveni Europeu de Drets Humans, en considerar que vulnerava el dret a la vida privada.

1.2.4. Opacitat algorítmica i dificultat de detecció

Els riscos discriminatoris que s'acaben d'analitzar resulten especialment difícils de detectar i controlar a causa de l'opacitat que caracteritza molts dels sistemes algorítmics, especialment aquells basats en tècniques complexes d'aprenentatge automàtic.

Tal com s'ha anat exposant al llarg del treball, la discriminació algorítmica acostuma a manifestar-se mitjançant formes indirectes de discriminació. En aquest sentit, la manca de transparència en els sistemes algorítmics dificulta la comprensió sobre quins factors han influït en les decisions adoptades per aquests i, en conseqüència, complica enormement la detecció, supervisió i impugnació dels possibles resultats discriminatoris.

A la pràctica, aquesta problemàtica es pot veure manifestada en alguns dels exemples que s'han utilitzat en els apartats anteriors per explicar els riscos discriminatoris. En el cas del sistema COMPAS, una de les principals dificultats va ser la manca d'accés al seu funcionament ja que, tot i no ser estrictament un sistema *blackbox*, "l'empresa propietària, es va negar a divulgar-ne el funcionament al·legant secret comercial. Això va impedir als acusats comprendre els criteris que sustentaven la seva qualificació de risc, afectant el dret al degut procés protegit per la Constitució dels EUA" (Lazcoz Moratinos i Castillo Parrilla, 2020; traducció pròpia).

En conseqüència, tot i les evidències existents sobre els possibles biaixos racials del sistema, resultava especialment complex determinar quines variables influïen realment en els resultats generats pel sistema i de quina manera eren ponderades aquestes variables. Aquesta manca de transparència va dificultar notablement la possibilitat de qüestionar i impugnar adequadament el funcionament del sistema COMPAS durant el cas *Loomis*.

Aquesta preocupació relativa a la falta de garanties suficients per controlar adequadament els riscos derivats dels sistemes algorítmics també es va posar de manifest en el cas neerlandès del sistema SyRI. En la mateixa resolució judicial oficial, dictada pel Tribunal del Districte de la Haia, es sosté que:

la legislació relativa a l'aplicació de SyRI no és prou transparent i verificable. Per aquest motiu, el tribunal declara en aquesta sentència que l'article 65 de la Llei SUWI i el capítol 5a del Decret SUWI no tenen efecte

vinculant, ja que són contradictoris amb l'article 8, apartat 2, del CEDH" (Tribunal del Districte de la Haia, 2020; traducció pròpia)¹⁴.

Per tant, aquests exemples evidencien que la transparència algorítmica no constitueix únicament una exigència tècnica, sinó un requisit indispensable per garantir un control efectiu sobre els riscos discriminatoris associats als sistemes d'IA d'alt risc. És a dir, sense un grau suficient de transparència, documentació i traçabilitat, resulta molt difícil detectar i impugnar possibles resultats discriminatoris generats pels sistemes algorítmics.

2. Els requisits dels sistemes d'IA d'alt risc com a mecanismes de prevenció de la discriminació

Tot i la rellevància dels riscos discriminatoris associats als sistemes d'IA que s'han exposat en l'apartat anterior, l'*AI Act* no estableix una definició precisa de discriminació algorítmica, sinó que les referències als possibles biaixos i efectes discriminatoris es troben disperses al llarg del reglament, especialment en algun considerant i en els requisits aplicables als sistemes d'alt risc.

No obstant això, el reglament sí que estableix un conjunt de garanties orientades a reduir aquests riscos. En aquest sentit, l'*AI Act* articula una estructura preventiva que pretén limitar els riscos discriminatoris dels sistemes algorítmics abans que aquests es materialitzin en decisions que afectin els drets fonamentals de les persones.

2.1. Governança de dades i prevenció de biaixos

En relació al primer risc discriminatori que s'ha explicat en l'apartat anterior –relatiu als biaixos preexistents que arrossegueu les dades d'entrada–, l'*AI Act* incorpora diversos requisits per garantir una governança adequada de les dades pel que fa als sistemes d'alt risc.

Cal destacar que el mateix reglament reconeix expressament la possibilitat que els sistemes algorítmics reproduïxin biaixos discriminatoris presents en les dades que utilitzen. Això es reflecteix en el considerant 67 del reglament, el qual remet expressament a l'article 10 del mateix reglament on es desenvolupen normativament les consideracions exposades al considerant mencionat, especialment en el 10.2 i el 10.3.

D'una banda, l'article 10.2 estableix que els sistemes d'IA d'alt risc han d'estar sotmesos a pràctiques adequades de governança, incloent

“l'examen a la vista de possibles biaixos que puguin afectar la salut i la seguretat de les persones, repercutir negativament en els drets fonamentals o donar lloc a discriminacions prohibides pel dret de la Unió, especialment quan els resultats de les dades influeixin en els inputs per a operacions futures¹⁵” (Art.10.2 del reglament (UE) 2024/1689; traducció pròpia).

¹⁴ La LLei SUWI, en concret els articles 64 i 65, eren la base jurídica del sistema SyRI i el Decret SUWI, en concret el capítol 5a, regulava el funcionament del sistema (Vegeu l'apartat de legislació).

¹⁵ És a dir, que també té en compte mesures per fer front al risc de retroalimentació exposat en l'apartat [1.2.3.](#) d'aquest capítol.

A més, el mateix apartat de l'article, exigeix l'adopció de mesures adequades per detectar, prevenir i mitigar possibles biaixos, així com identificar possibles llacunes o deficiències en les dades utilitzades pels sistemes.

D'altra banda, el següent apartat del mateix article (10.3) estableix que les dades d'entrenament, validació i prova dels sistemes d'alt risc han de ser pertinents, suficientment representatives i, en la mesura del possible, lliures d'errors i completes tenint en compte les finalitats previstes. Aquestes exigències adquireixen una rellevància especial en relació amb els riscos discriminatoris derivats de dades històriques esbiaixades, com s'ha exemplificat de manera clara en l'apartat [1.2.1](#) amb el cas del sistema de contractació d'Amazon.

En la mateixa línia, resulta molt rellevant destacar el contingut de l'article 10.5, el qual preveu la possibilitat que –de manera excepcional i sota estrictes garanties– els proveïdors dels sistemes d'IA d'alt risc puguin tractar categories especials de dades personals quan sigui estrictament necessari per detectar i corregir possibles biaixos discriminatoris dels sistemes que proveeixen, sempre de conformitat amb el règim europeu de protecció de dades¹⁶ que es menciona en l'article.

Aquesta previsió esdevé especialment sensible des de la perspectiva dels drets i les llibertats fonamentals, ja que les categories de dades especials a les quals es fa referència, són aquelles relatives a l'origen racial o ètnic, les conviccions religioses, la salut o la discapacitat, entre altres. Per aquest motiu, l'article condiciona aquest tractament al compliment d'un conjunt ampli d'estrictes garanties¹⁷.

D'aquesta manera, el reglament reconeix implícitament que la detecció efectiva de determinades formes de discriminació algorítmica, a vegades, requereix el tractament de les mateixes categories que es pretenen protegir d'un tracte discriminatori. Tot i això, com s'acaba d'esmentar, conscient dels riscos que comporta aquest tractament, el reglament l'envolta d'un conjunt de garanties destinades a preservar els drets i les llibertats fonamentals.

¹⁶ Reglament (UE) 2016/679, Reglament (UE) 2018/1725 i Directiva (UE) 2016/680.

¹⁷ Les garanties previstes en l'article 10.5 del reglament (UE) 2024/1689 són (traducció pròpia):

- la detecció i correcció de biaixos no es pot fer eficaçment mitjançant el tractament d'altres dades, incloses les dades sintètiques o anonimitzades;
- les categories especials de dades personals estiguin subjectes a limitacions tècniques quant a la reutilització de les dades personals, ia mesures de seguretat i de preservació de la intimitat d'última generació, inclosa la pseudonimització;
- les categories especials de dades personals estiguin subjectes a mesures que garanteixin que les dades personals tractades estiguin segures, protegides, subjectes a les salvaguardes adequades, inclosos controls estrictes i documentació de l'accés, per evitar usos indeguts i garantir que només les persones autoritzades tinguin accés a aquestes dades personals amb les obligacions de confidencialitat adequades;
- les categories especials de dades personals no han de ser transmeses, transferides o accessibles altrament a altres parts;
- les categories especials de dades personals se suprimen una vegada que s'hagi corregit el biaix o les dades personals hagin arribat al final del seu període de conservació, cosa que passi primer;
- els registres de les activitats de tractament de conformitat amb els Reglaments (UE) 2016/679 i (UE) 2018/1725 i la Directiva (UE) 2016/680 inclouen les raons per les quals el tractament de categories especials de dades personals era estrictament necessari per detectar i corregir biaixos, i per què aquest objectiu no podia assolir aquest objectiu.

Per tant, mitjançant aquests requisits relatius a la qualitat i la governança de les dades, l'*AI Act* introdueix mecanismes preventius orientats a reduir el risc dels sistemes d'IA d'alt risc de que reproduïxin automàticament desigualtats estructurals presents en les dades d'entrenament, validació i prova.

2.2. Transparència, documentació i traçabilitat

Tanmateix, la detecció i correcció dels biaixos discriminatoris no depèn únicament de la qualitat de les dades utilitzades, sinó també del fet de poder comprendre, documentar i rastrejar el funcionament dels sistemes algorítmics. És per això que, juntament amb els requisits vinculats a la qualitat i governança de dades, l'*AI Act* incorpora també diverses exigències de transparència, documentació i traçabilitat orientades a reduir l'opacitat que caracteritza aquests sistemes.

La rellevància d'aquests requisits es posa de manifest en els considerants 66, 71 i 72, en els que es destaca la importància de la necessitat de que els sistemes d'IA d'alt risc estiguin subjectes a determinats requisits de documentació, registre i transparència, que facilitin la supervisió dels sistemes i garanteixin la traçabilitat dels resultats generats. Aquesta lògica es troba regulada principalment en els articles 11, 12 i 13 del reglament.

En primer lloc, l'article 11 de l'*AI Act* estableix que la documentació tècnica d'un sistema d'IA d'alt risc s'ha d'elaborar abans de la seva comercialització o posada en servei i, aquesta informació, ha de poder demostrar que:

el sistema d'IA d'alt risc compleix els requisits que estableix aquesta secció i proporcioni a les autoritats nacionals competents i als organismes notificats la informació necessària de manera clara i completa per avaluar la conformitat del sistema d'IA amb aquests requisits. Contindrà, com a mínim, els elements establerts a l'Annex IV¹⁸. (Art.11.1 del reglament (UE) 2024/1689; traducció pròpia)

La importància d'aquest requisit resideix en el fet que permet generar evidències verificables sobre els criteris seguits durant la vida útil del sistema, cosa que facilita els processos d'eventuals auditories, supervisions o controls posteriors.

En relació a la traçabilitat, l'article 12 incorpora requisits relatius al fet que els sistemes d'IA d'alt risc han de permetre tècnicament el registre automàtic d'esdeveniments (logs) durant la vida útil del sistema. Aquests registres funcionen mitjançant l'enregistrament automàtic d'esdeveniments rellevants durant el desenvolupament del sistema, permetent reconstruir posteriorment determinades operacions o decisions generades pel mateix.

Aquesta traçabilitat del funcionament dels sistemes d'alt risc no només ha de servir de suport per facilitar la reconstrucció de determinades decisions, sinó que també, segons l'article 12.1, ha de permetre identificar situacions que puguin afectar la salut, la seguretat o els drets fonamentals de les persones (art. 79.1 del reglament (UE) 2024/1689), facilitar el seguiment posterior a la comercialització del sistema (art. 72 del reglament (UE) 2024/1689) i reforçar

¹⁸ Aquest mínim de documentació tècnica establert a l'Annex IV fa referència a informació sobre el disseny i funcionament del sistema, les seves finalitats previstes, les dades utilitzades, els processos de desenvolupament i validació, les mesures de gestió de riscos, els mecanismes de supervisió humana i, entre altres, els elements necessaris per avaluar el compliment dels requisits establerts pel reglament.

els mecanismes de supervisió previstos en reglament (art. 26.5 del reglament (UE) 2024/1689).

Aquesta exigència resulta especialment rellevant des de la perspectiva de la no-discriminació ja que permet facilitar la identificació de patrons discriminatoris i contribueix a determinar si certs resultats poden estar vinculats a biaixos presents en el sistema.

D'altra banda, l'article 13 preveu uns requisits específics de transparència destinats a garantir que els desplegadors dels sistemes d'IA d'alt risc puguin interpretar i utilitzar adequadament els resultats generats pel sistema i, a més, garantir que tant els proveïdors com els desplegadors compleixin amb les obligacions previstes a la secció 3 del Capítol III –les quals deriven dels requisits aplicables als sistemes d'IA d'alt risc que s'estan analitzant en el present apartat que, com s'ha mencionat a la introducció, estan previstos a la secció 2 del Capítol III–.

Els requisits de transparència previstos en aquest article fan referència a que el disseny i el desenvolupament d'aquests sistemes han de garantir que el seu funcionament sigui suficientment transparent. Alhora, els sistemes han d'anar acompanyats d'un mínim d'informació d'ús que, segons l'article 13.2, ha de ser concisa, completa, correcta, clara i, pertinent, accessible i comprensible per als desplegadors.

És a dir, aquesta informació resulta imprescindible perquè els responsables de la utilització d'aquests sistemes puguin interpretar correctament els resultats generats i, així, poder detectar possibles anomalies o conductes discriminatòries.

Per tant, com s'ha anat esmentant al llarg d'aquest treball, la transparència en el marc de l'*AI Act* no s'ha d'entendre com un mecanisme d'accés a la informació, sinó com una eina orientada a fer front a l'opacitat algorítmica mitjançant la documentació i la traçabilitat.

Tanmateix, tot i el gran pas que suposa l'*AI Act* en matèria de regulació harmonitzada de la IA i, en el que respecta aquest treball, la rellevància del paper dels requisits analitzats com a mecanisme de prevenció de discriminació algorítmica, a la pràctica encara presenten algunes limitacions. En aquest sentit, que hi hagi disponibilitat de documentació o registres sobre el sistema no implica que les persones afectades pels resultats generats puguin comprendre efectivament els criteris que han conduït a un determinat resultat.

En conseqüència, encara que aquests mecanismes contribueixin a reduir l'opacitat algorítmica, no sempre garanteixen que les decisions adoptades automàticament per la IA –o amb el seu suport– siguin plenament explicables, justificables o fàcilment impugnables, especialment quan afecten els drets i oportunitats de les persones. En aquests casos tan sensibles, no només és necessari que les decisions adoptades puguin ser adequadament justificades, revisades i, si és necessari, impugnades, sinó que també han d'existir mecanismes efectius de rendició de comptes que permetin atribuir responsabilitats pels resultats produïts pels sistemes d'IA.

En aquest context, s'ha advertit del “difícil equilibri entre el caràcter no discriminatori i la transparència dels sistemes de IA vaticina una futura tendència regressiva i menys garantista

cap a la protecció dels drets humans a la UE, especialment per a certes persones o grups vulnerables” (EDRi, 2022; Avello, 2023; Garrido Carillo, 2023; citat per La Espina, 2024; traducció pròpia).

D’aquesta manera, convé assenyalar que la complexitat tècnica d’aquests sistemes, la participació de múltiples actors en el seu desenvolupament i desplegament, així com les dificultats associades a l’atribució de responsabilitats, poden limitar l’efectivitat pràctica dels mecanismes tradicionals de control. Aquestes dificultats posen de manifest les tensions existents a l’hora de garantir, a la vegada, el desenvolupament i la innovació tecnològica, la transparència dels sistemes i la protecció efectiva dels drets fonamentals.

És per això que, encara que la documentació, la traçabilitat i la transparència siguin eines fonamentals per reduir l’opacitat algorítmica, la seva utilitat real depèn de l’existència de mecanismes capaços de transformar aquesta informació en formes efectives de control.

Aquesta necessitat explica que l’*AI Act* complementi els requisits previstos en els articles 10, 11, 12 i 13 amb requisits específics de supervisió humana, orientats a garantir que les persones mantinguin una capacitat real d’intervenció i control sobre el funcionament dels sistemes d’IA d’alt risc.

2.3. Supervisió humana

En relació amb aquesta qüestió, el considerant 73 destaca la necessitat de que els sistemes d’IA d’alt risc es dissenyin i es desenvolupin de manera que les persones físiques puguin supervisar-ne el funcionament, garantir que s’utilitzen segons allò previst i que els seus impactes s’aborden al llarg del cicle de vida del sistema.

Així mateix, el considerant subratlla la importància que les persones encarregades de fer la supervisió disposin de la competència, la formació i l’autoritat necessàries per exercir aquesta funció, així com que els sistemes incorporin mecanismes que facilitin la intervenció humana quan es requereixi. Aquestes consideracions es desenvolupen normativament a l’article 14 del reglament, el qual estableix els requisits concrets de supervisió humana aplicables als sistemes d’IA d’alt risc.

La idea-força tant del considerant 73 com de l’article 14 del reglament, gira entorn a la necessitat de supervisió humana efectiva, de manera que s’eviti una dependència cega de decisions automatitzades (García-Llorente i Olmeda, 2025; traducció pròpia). Això es pot veure reflectit en el segon apartat de l’article 14, el qual estableix que la supervisió humana ha de tenir com a finalitat prevenir o reduir al mínim els riscos que els sistemes d’IA d’alt risc puguin generar per a la salut, la seguretat o els drets fonamentals de les persones. D’aquesta manera, el reglament parteix de la idea que els resultats generats pels sistemes d’IA no han de ser acceptats automàticament, sinó que han d’estar sota control i intervenció humana.

Aquest requisit adquireix una rellevància especial des de la perspectiva de la no-discriminació ja que, com s’ha analitzat al llarg del treball, els resultats discriminatoris dels sistemes d’IA poden tenir diferents causes. Davant aquest fet, la supervisió humana es configura com un mecanisme complementari als requisits analitzats anteriorment, amb la

finalitat d'actuar en aquells supòsits en què, malgrat l'aplicació de les garanties relatives a la governança de dades, la transparència, la documentació i la traçabilitat, continuï existint el risc que es generin resultats discriminatoris.

A més, l'article 14.3 preveu que les persones encarregades d'aquesta tasca disposin de la capacitat necessària per intervenir en el funcionament del sistema quan sigui necessari. D'una banda, entre les facultats previstes en aquest apartat de l'article, destaca la capacitat de poder –i saber– “decidir, en qualsevol situació concreta, no utilitzar el sistema d'IA d'alt risc o fer cas omís, anul·lar o invertir el resultat del sistema d'IA d'alt risc (art. 14.3 d) del reglament (UE) 2024/1689). D'altra banda, destaca la capacitat dels encarregats de la supervisió de “ser conscients de la possible tendència a confiar automàticament o en excés en els resultats produïts per un sistema d'IA d'alt risc (biaix d'automatització)¹⁹” (art. 14.3 b) del reglament (UE) 2024/1689).

Amb tot això, el que es pretén és evitar que els sistemes d'IA d'alt risc operin de manera autònoma en els àmbits previstos en l'Annex III, garantint una capacitat real d'intervenció i control humà sobre els resultats generats. Com molt bé argumenta Hernández (2024; traducció pròpia):

La supervisió humana permet corregir possibles biaixos i errors, assegurar la transparència en els processos i proporcionar una capa addicional de control que eviti abusos i vulneracions de la privadesa (...) A més, la supervisió humana facilita la rendició de comptes, permetent que les decisions preses per la IA siguin revisades i, si escau, rectificades per persones capacitades mitjançant auditories periòdiques i la creació de comitès ètics, assegurant que operin de manera justa, transparent i equitativa.

Malgrat tots els avantatges que suposa aquest requisit en la teoria, l'existència d'un mecanisme de supervisió humana, en la pràctica, no garanteix un control efectiu sobre les decisions adoptades per la IA o amb el seu suport.

Precisament, el mateix biaix d'automatització que l'article 14.3 pretén prevenir, posa de manifest una de les principals limitacions pràctiques d'aquest requisit. Es pot donar el cas que les persones responsables de la tasca de supervisió, desenvolupin una confiança excessiva respecte els resultats generats pel sistema, provocant que la supervisió humana es converteixi en una pràctica banal, limitada a validar automàticament les recomanacions o decisions del sistema sense que es dugui a terme una revisió crítica real dels resultats.

Per tant, l'efectivitat de la supervisió humana dependrà, no només de la seva previsió en el reglament, sinó també de la formació, competències i capacitat real d'intervenció de les persones encarregades d'exercir-la. És a dir, la supervisió humana només podrà esdevenir una garantia efectiva davant els riscos discriminatoris dels sistemes d'IA quan la intervenció humana sigui una intervenció realment informada, crítica i capaç de qüestionar els resultats generats pel sistema.

¹⁹ En particular en el cas dels sistemes d'IA d'alt risc utilitzats per proporcionar informació o recomanacions per a la presa de decisions per part de persones físiques.

Conclusions

La progressiva introducció dels sistemes d'IA en els processos de presa de decisions ha incrementat la preocupació pels possibles resultats discriminatoris que aquests sistemes poden generar, sobretot, quan operen en àmbits especialment sensibles per als drets i les oportunitats de les persones.

Davant aquest escenari, la Unió Europea ha fet un pas significatiu en matèria de regulació de la IA amb l'aprovació de l'*AI Act*, un marc jurídic harmonitzat que reflecteix la voluntat de conciliar la innovació i desenvolupament tecnològic amb la protecció dels drets fonamentals i dels valors sobre els quals es fonamenta la Unió. En aquest context, els sistemes d'IA d'alt risc adquireixen una posició central dins el reglament, ja que són aquells que operen en els àmbits especialment sensibles que s'acaben de mencionar i que, per tant, presenten una major capacitat potencial d'incidir negativament sobre els drets i oportunitats de les persones.

És precisament en relació amb aquesta categoria de sistemes que s'ha plantejat la pregunta que ha orientat el present treball, relativa a fins a quin punt els requisits previstos en l'*AI Act* aplicables a aquests sistemes contribueixen a afrontar els riscos discriminatoris que hi estan associats.

Una de les principals amenaces de fons que s'ha observat és que la discriminació algorítmica es manifesta majoritàriament en forma de discriminació indirecta, fet que comporta que l'aparent neutralitat que poden presentar els resultats generats pels sistemes d'IA d'alt risc no garanteixi la seva neutralitat real. A aquesta problemàtica se li suma l'opacitat que caracteritza a aquests sistemes, la qual augmenta més encara la dificultat de detectar els possibles riscos de discriminació abans de que es materialitzin.

Davant d'aquesta amenaça, la UE va dissenyar el Reglament (UE) 2024/1689 des d'una lògica preventiva i basada en el nivell de risc. D'aquesta manera, excepte els que estan catalogats com a risc inacceptable, més que pretendre eliminar completament els riscos, el reglament estableix un conjunt de requisits ex-ante orientats a reduir-los i gestionar-los, això sí, dins uns límits considerats acceptables. Per tant, havent analitzat aquests requisits, el primer que s'ha pogut concloure és que per fer front als riscos de discriminació associats als sistemes d'IA d'alt risc, no és suficient amb un únic mecanisme preventiu, sinó que es requereix la combinació d'un conjunt de garanties complementàries entre si.

Així doncs, els requisits analitzats exerceixen funcions diferenciades però complementàries: la governança de dades busca reduir els biaixos en els conjunts de dades utilitzats pels sistemes; la documentació, la traçabilitat i la transparència lluiten contra l'opacitat algorítmica, facilitant la comprensió, la supervisió i el control del funcionament dels sistemes; i la supervisió humana, com molt bé s'ha vist amb la cita d'Hernández (2024; traducció pròpia), permet corregir possibles biaixos i errors, assegurar la transparència en els processos i proporcionar una capa addicional de control que eviti abusos i vulneracions de la privadesa.

A més, es pot concloure que, un cop analitzats i d'acord amb una gran part de la literatura existent, els requisits previstos contribueixen de manera significativament positiva a la lluita contra la discriminació associada als resultats generats pels sistemes d'IA d'alt risc i, per tant, a protegir el principi d'igualtat com a valor fonamental de la UE.

Ara bé, aquesta contribució no és absoluta, en primer lloc perquè com s'acaba d'esmentar el principal objectiu del reglament no és eliminar per complet aquests riscos, sinó a prevenir-los i limitar-los. També, durant la investigació, s'ha pogut comprovar que la disponibilitat de la informació sobre el funcionament de determinats sistemes, no garanteix la comprensió dels resultats generats per part de les persones afectades.

També, com s'ha assenyalat a l'apartat anterior, que existeixi un mecanisme de supervisió humana no assegura completament el control efectiu sobre les decisions generades per la IA –o amb el seu suport–. Per últim, també s'ha de tenir en compte el fet que les tecnologies d'IA evolucionen exponencialment, cosa que presenta un repte addicional a la contribució a fer front a la discriminació, ja que aquests requisits hauran de ser capaços d'adaptar-se al ritme de les transformacions tecnològiques, el qual és molt superior a l'evolució normativa.

Cal destacar que les conclusions extretes s'han de tenir en compte considerant les limitacions que deriven de la metodologia seguida per la realització del present treball. En primer lloc, l'estudi s'ha realitzat des d'una perspectiva teòrico-normativa, de manera que no s'ha abordat empíricament l'aplicació pràctica ni l'efectivitat real dels requisits analitzats. A més, com que l'aprovació d'aquest reglament és molt recent, no ha passat el temps suficient per poder valorar l'impacte real d'aquest requisits en relació a la prevenció dels riscos discriminatoris associats als sistemes d'IA d'alt risc.

Aquesta última limitació obre interrogants importants i difícils de resoldre en l'estat actual de desenvolupament del reglament. En particular, resultaria interessant analitzar l'evolució d'aquests requisits davant l'extraordinàriament ràpida evolució dels sistemes d'IA i dels seus usos per comprovar la seva efectivitat real.

Legislació

- Constitución Española. (1978). Boletín Oficial del Estado. <https://www.boe.es/buscar/act.php?id=BOE-A-1978-31229>
- European Commission. (2021). Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Fostering a European approach to artificial intelligence (COM(2021) 205 final). EUR-Lex. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021DC0205>
- European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (COM(2021) 206 final).

https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

- Països Baixos. (2001). Besluit SUWI. Overheid.nl. [Overheid.nl - Besluit SUWI](#)
- Països Baixos. (2001). Wet structuur uitvoeringsorganisatie werk en inkomen (Wet SUWI). Overheid.nl. [Overheid.nl - Wet SUWI](#)
- Parlamento Europeo y Consejo de la Unión Europea. (2024). 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 (Reglamento de Inteligencia Artificial). Diario Oficial de la Unión Europea, L 1689. <https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32024R1689>
- Unión Europea. (2010). Carta de los Derechos Fundamentales de la Unión Europea. Diario Oficial de la Unión Europea, C 83, 389-403. <https://www.boe.es/doue/2010/083/Z00389-00403.pdf>
- Unión Europea. (2010). Versión consolidada del Tratado de la Unión Europea, Maastricht, 7-2-1992. Diario Oficial de la Unión Europea, 30-3-2010, C 83, 13-47. <https://www.boe.es/doue/2010/083/Z00013-00046.pdf>
- Unión Europea. (2006). *Directiva 2006/54/CE del Parlamento Europeo y del Consejo, de 5 de julio de 2006, relativa a la aplicación del principio de igualdad de oportunidades e igualdad de trato entre hombres y mujeres en asuntos de empleo y ocupación (refundición)*. Diario Oficial de la Unión Europea, L 204, 23-36. <https://www.boe.es/doue/2006/204/L00023-00036.pdf>

Jurisprudència

- Tribunal de Justicia de la Unión Europea. (2026, 21 d'abril). *Assumpte C-769/22 Comissió Europea c. Hongria*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62022CJ0769>
- Tribunal del Districte de La Haia. (2020, 5 de febrer). *C/09/550982 / HA ZA 18-388. (ECLI:NL:RBDHA:2020:1878)*. Rechtspraak. <https://uitspraken.rechtspraak.nl/details?id=ECLI:NL:RBDHA:2020:1878>

Bibliografia

- Arochena, J. F. (2024). Inteligencia artificial y sesgos discriminatorios: ¿es necesario un nuevo concepto de discriminación algorítmica?. *IgualdadES*, 11, 97, 123, 26. <https://www.cepc.gob.es/sites/default/files/2024-12/40676igd1104lousada-arochena.pdf>
- Banchio, P. (2025). Reglamento de la Unión Europea sobre Inteligencia Artificial. <https://amicuscuriaequorum.wordpress.com/wp-content/uploads/2025/07/volume-14-2025.pdf>

- Bandy, J. (2021). Problematic machine behavior: A systematic literature review of algorithm audits. *Proceedings of the acm on human-computer interaction*, 5 (CSCW1), 1-34. <https://dl.acm.org/doi/abs/10.1145/3449148>
- Dastin, J. (2018, October 10). *Insight - Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters. <https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>
- de Zárate Alcarazo, L. O. (2023). Sesgos de género en la inteligencia artificial. *Revista de occidente*, 502. https://ortegaygasset.edu/wp-content/uploads/2023/03/RevistadeOccidente_Marzo2023_L.Ortiz_de_Zarate.pdf
- García-Llorente, C., & Olmeda, I. (2025). Opacidad algorítmica y Derechos Humanos: Comparación crítica UE-EEUU. In *Actas del VIII Congreso sobre Derechos Humanos Derechos Humanos e Inteligencia Artificial*, 265-288. <https://derechoshumanos.maiel.org/wp-content/uploads/2026/01/Actas-VIII-Congreso-Derechos-Humanos-e-Inteligencia-Artificial.pdf#page=265>
- Hernández, J. S. (2024). La necesaria evolución de los derechos humanos: encrucijada digital y discriminación algorítmica. *Revista de Derecho UNED*, (33), 617-637. https://media.proquest.com/media/hms/PFT/1/hi4kZ?_s=yIDZSfoE%2F%2FZ28qCSiCj1mKdKcVc%3D
- Iturmendi Rubia, J. M. (2023). Algorithmic discrimination and its impact on human dignity and human rights. Special reference to immigrants. *Deusto Journal of Human Rights*, 12, 257-284. <https://djhr.revistas.deusto.es/article/view/2910/3510>
- La Spina, E. (2024). La regulación europea de la IA ante los sesgos y riesgos de discriminación algorítmica en contextos migratorios - The European regulation on AI in the face of the biases and risks of algorithmic discrimination in migration contexts. *Revista CIDOB d'Afers Internacionals*, (138), 171–194. <https://www.jstor.org/stable/27393914>
- Lazcoz Moratinos, G., & Castillo Parrilla, J. A. (2020). Valoración algorítmica ante los derechos humanos y el Reglamento General de Protección de Datos: el caso SyRI. *Revista chilena de derecho y tecnología*, 9(1), 207-225. https://www.scielo.cl/scielo.php?pid=S0719-25842020000100207&script=sci_arttext&tln_g=en
- Martínez-Pujalte, C. C. (2005). Los principios constitucionales de igualdad de trato y de prohibición de la discriminación: un intento de delimitación. *Cuadernos Constitucionales de la Cátedra Fadrique Furió Ceriol*, (50), 193-218.
- Pandey, D., Niwaria, K., & Chourasia, B. (2019). Machine learning algorithms: a review. *Mach. Learn*, 6(2), 916-922. <https://n9.cl/4nryj>

- Rojas, J. (2024). *De algoritmos y derechos: reflexiones sobre la privacidad y la protección de datos en un mundo interconectado por la IA*. Noticias Jurídicas. <https://noticias.juridicas.com/conocimiento/tribunas/18728-de-algoritmos-y-derechos:-reflexiones-sobre-la-privacidad-y-la-proteccion-de-datos-en-un-mundo-interconectado-por-la-ia/>
- Rouhiainen, L. (2018). *Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro*. Alienta Editorial.
- Sánchez Martínez, M. O. (2025). Brechas digitales y sesgos en la inteligencia artificial: ¿una nueva forma de discriminación?. <https://repositorio.unican.es/xmlui/bitstream/handle/10902/36979/BrechasDigitalesySesgo.sIA.pdf?sequence=3&isAllowed=y>
- Sierra Hernáiz, E. (2018). La discriminación de género en la relación laboral: igualdad formal, igualdad material y acción positiva. *Foro: Revista de Derecho*, 49-64. https://scielo.senescyt.gob.ec/scielo.php?pid=S2631-24842018000200049&script=sci_abstract&tlng=e
- Trail ML. (13 de febrer de 2026). *EU AI Act: How risk is classified*. <https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified>