Per una IA responsable: Principis, Indicadors i Observables

Model PIO

fefic #SomFires

Generalitat de Catalunya

OEIAC

Universitat de Girona

L'OEIAC dins l'estratègia CATALONIA.AI

La Generalitat de Catalunya impulsa l'Estratègia d'Intel·ligència Artificial (IA) de Catalunya que, amb el nom de **CATALONIA.AI**, realitzarà un programa d'actuacions per enfortir l'ecosistema català en IA.

L'Estratègia CATALONIA.AI coincideix amb el **"Pla Estratègic UdG2030: la Suma d'Intel·ligències"** de la Universitat de Girona.

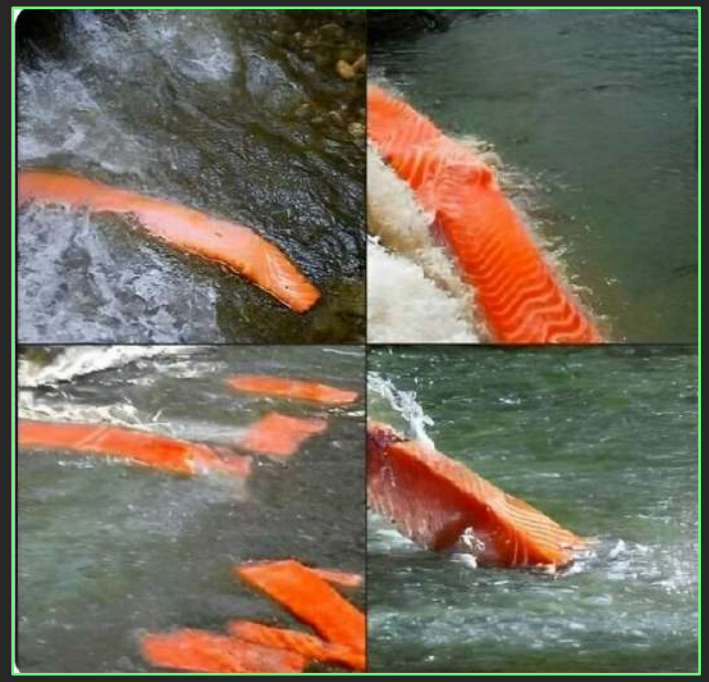
Estudiar les ***conseqüències*** ètiques, socials i legals, així com els riscos i oportunitats de la implantació de la IA en la vida diària a Catalunya, des d'una òptica plenament transversal.

Punt de partida

Sembla obvi però no ho és

- Els sistemes d'IA **no poden respondre preguntes** on les respostes no estiguin d'alguna forma en el seu conjunt de formació.
- Els sistemes d'IA **no poden resoldre problemes** sobre els quals no han estat entrenats (dades i arquitectura).
- Els sistemes d'IA **no poden adquirir noves habilitats** sense la nostra ajuda (recollida d'informació a l'etiquetatge i anàlisi).

Sembla obvi, però no ho és



Sistemes d'IA ètics i legals

- Ubiquen les persones al centre.
- Compleix amb la regulació aplicable.
- Respectuós amb els acords socials.
- Adaptabilitat del sistema.
- Compleix amb el principi de transparència, justícia, seguretat, responsabilitat, autonomia, privacitat i sostenibilitat.

Preguntes inicials

Quin problema intentem resoldre amb el sistema d'IA?

El necessitem ?

Com es resoldrà?

Quins beneficis aportarà?

El Model PIO

Què és el Model PIO?

1. Un model d'autoavaluació integrat per preguntes.

2. És fonamenta en la normativa aplicable i recomanacions ètiques i legals.

3. Ajuda les persones usuàries a identificar si el seu sistema compleix amb els principis legals i ètics.

Principis, Indicadors, Observables

P	I	O
Principis	**Indicadors**	**Observables**
7 principis	+130 preguntes	Resultats globals Matriu de riscos Punts de millora

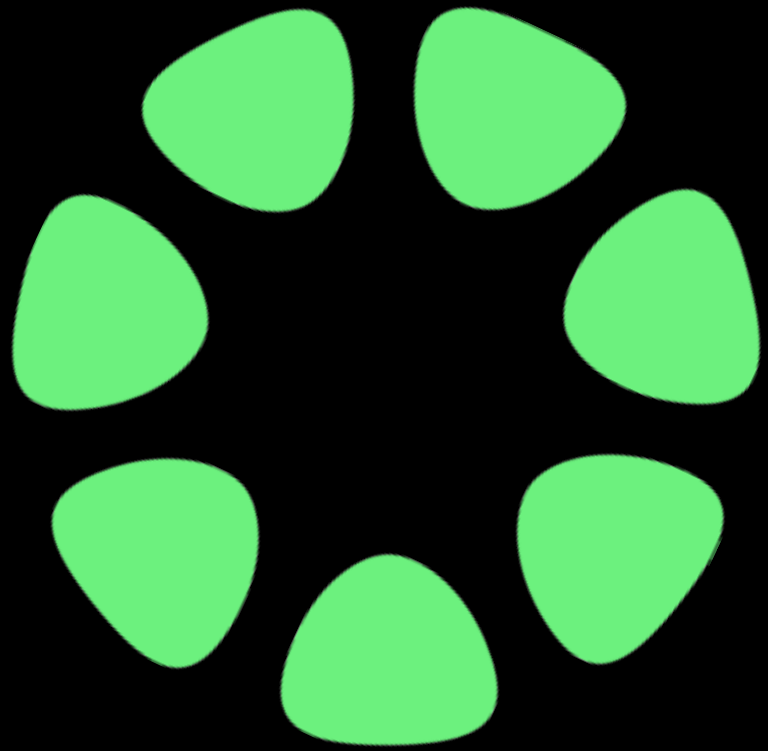
Els 7 principis

Sostenibilitat:
Impacte mediambiental i social així com governança.

Autonomia:
Control de l'usuari, llibertat d'elecció i consentiment.

Privacitat:
Protecció de dades, llibertat i confiança.

Seguretat i no maleficència:
Evitar danys previsibles, no intencionats i mecanismes de seguretat.

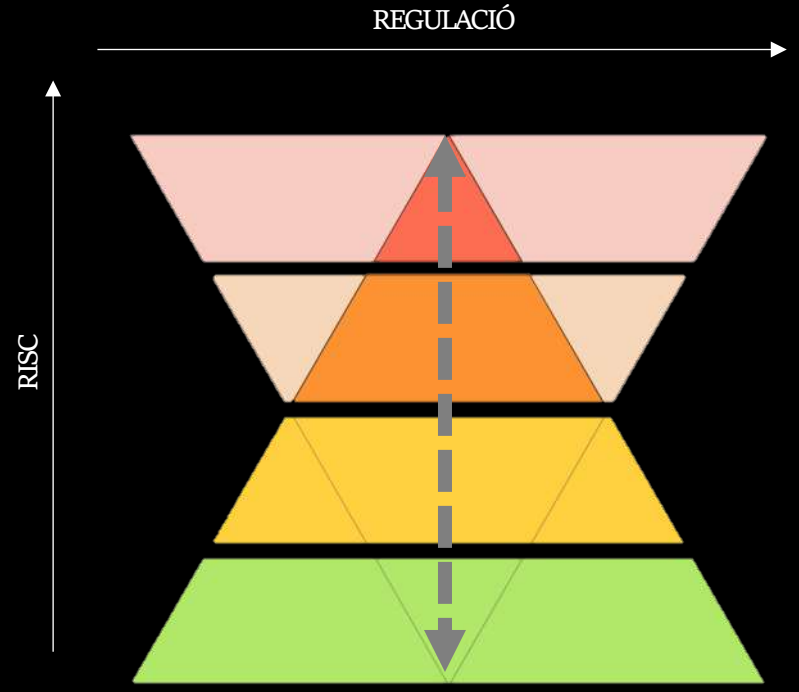


Transparència:
Explicació, interpretació i comunicació correcta.

Justícia i equitat:
Tracte just i imparcial, prevenció, seguiment i mitigació de biaixos i discriminacions.

Responsabilitat i retiment de comptes:
Integritat en les accions i decisions i atribució de responsabilitat.

Avaluació adaptada a les categories de risc



Risc Inacceptable
e.g. Puntuació social

Prohibits.

Alt risc
e.g. Classificació CVs

Permès amb obligacions estrictes i la realització d'una avaluació de compliment.

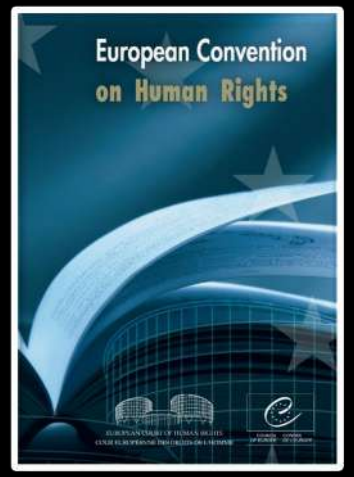
Risc limitat
e.g. Chatbots

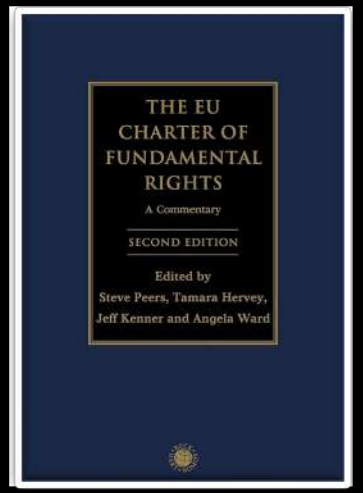
Permès amb obligacions de transparència associades.

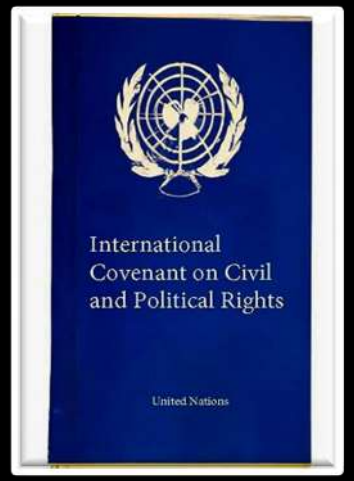
Risc mínim
e.g. Videojocs

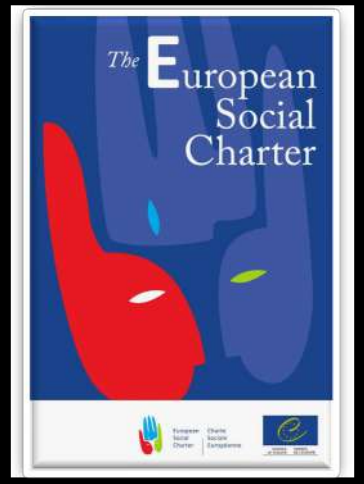
Permès amb restriccions mínimes i s'anima a seguir codis de conducta.

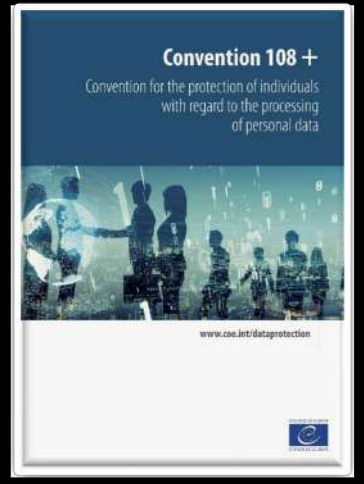
Perquè drets?











Perquè obligacions?

Convenció Europea de Drets Humans.

Pacte Internacional de Drets Civils i Polítics.

Declaració Universal de Drets Humans.

Carta Social Europea.

Carta Europea de Drets Fonamentals.

Reglament general de protecció de dades.

Carta de drets digitals.

Constitució Espanyola.

Estatut d'Autonomia de Catalunya.

RIA

Convenció 108+
Per a la protecció de les persones respecte el tractament automatitzat de dades de caràcter personal

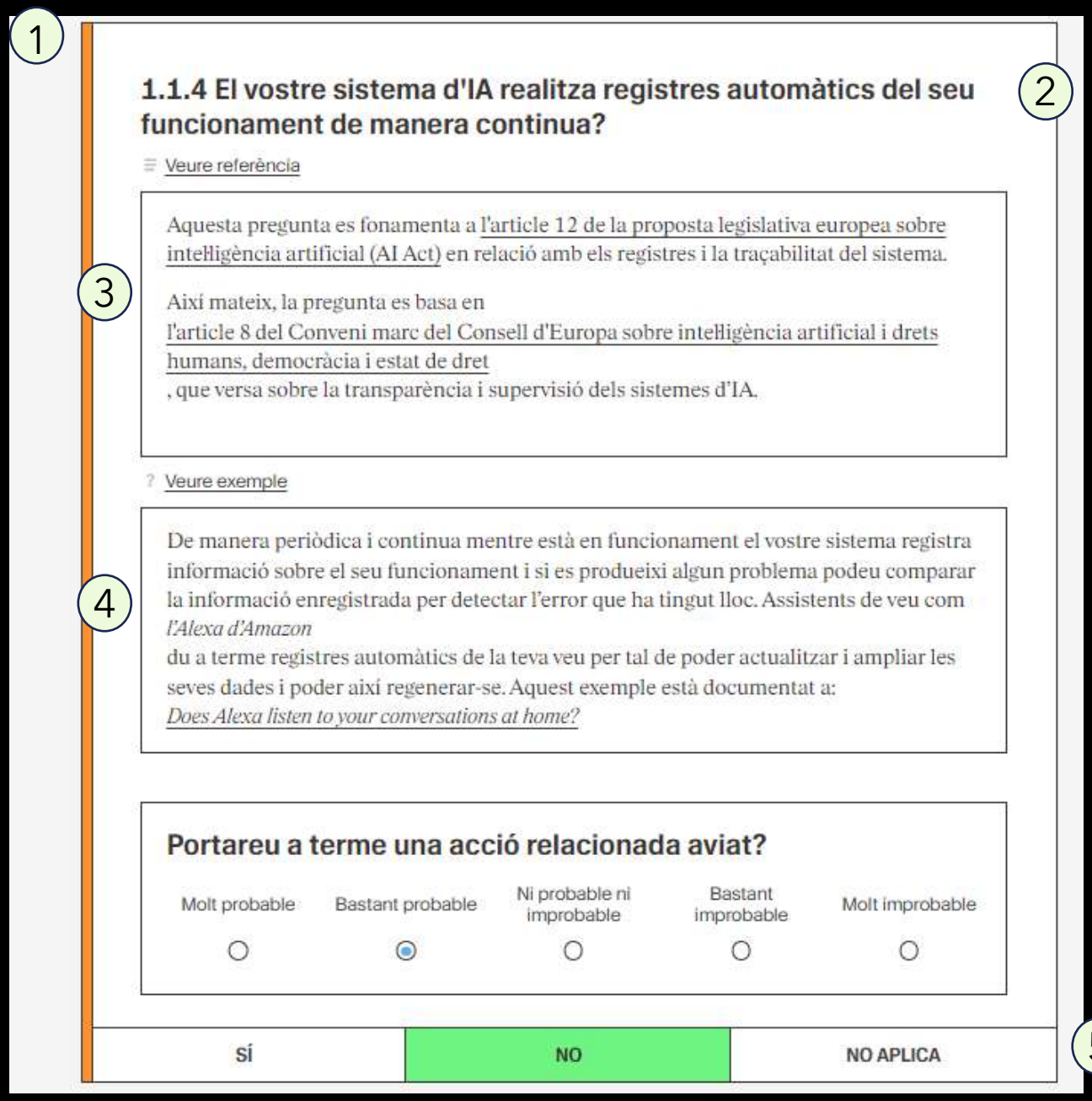
Directiva 2001/29/CE.

Directiva 2009/24/CE.

Directiva 2019/790/CE.

Funcionament


1. Preavaluació 1
2. Preavaluació 2
3. Resultats de la preavaluació
4. Inici de l'avaluació
5. Recollida de dades
6. Obtenció de les mètriques
7. Visualització dels resultats



1- Franja de color.

2- Pregunta i numeral.

3- Referència legal.

4- Exemple.

5- Resposta.

2.2.2 Heu establert mecanismes que permetin mitigar els possibles biaixos de classificació en el vostre sistema d'IA, siguin per categories canviants o inadequades segons el context sociocultural?

Veure referència

Aquesta pregunta fa referència als articles 9, 10.3, 15.3, 53, 55.1 i 95 del Reglament 2024/1689 del Parlament i del Consell sobre intel·ligència artificial (AI Act), relatius a les mesures de gestió de riscos, la gestió de dades, la resiliència, robustesa dels sistemes d'IA, les obligacions de les persones proveïdores de sistemes de propòsit general i la promoció de codis de conducta d'aplicació voluntària en matèria de grups vulnerables.

Així mateix, la pregunta es basa en l'article 4.2 de la proposta de directiva del Parlament Europeu i del Consell relativa a normes de responsabilitat civil extracontractual a la intel·ligència artificial (AI Liability Act), relativa a la presumpció de culpa dels danys provocats per una IA.

A més, la pregunta es fonamenta en els articles 10, 12 i 16 del Conveni marc del Consell d'Europa sobre intel·ligència artificial i drets humans, democràcia i estat de dret, que versen sobre la igualtat i la no discriminació, la fiabilitat dels sistemes d'IA i el marc de gestió de riscos i impactes. Juntament amb el dret humà a la no discriminació, consagrat a l'article 14 del Conveni Europeu de Drets Humans, l'article 21 de la Carta de Drets Fonamentals de la UE, l'article 26 del Pacte Internacional de Drets Civils i Polítics i l'article 7 de la Declaració Universal dels Drets Humans.

La pregunta també està relacionada amb el principi *Self-identification* del document A Human Rights-Based Approach To Data de l'ACNUDH l'apartat de *Diversity, Non-discrimination and Fairness* d'ALTAI.

Veure exemple

Per tal de garantir la mitigació de biaixos es pot recopilar dades de manera equilibrada i una de les tècniques és el submostreig aleatori o *Random Undersampling*. Consisteix en reduir la mida de la classe de majoritària i així s'iguala amb la classe minoritària. En el següent enllaç hi trobareu informació:
https://www.sciencedirect.com/science/article/pii/S1877050919313456 i

https://ieeexplore.ieee.org/document/9078901.

SÍ	NO	NO APLICA

5.1.2 Heu establert mecanismes de protecció de dades per tal que les dades de caràcter personal subministrades per la persona interessada no siguin venudes, llogades ni cedides, sense el seu consentiment?

Veure referència

Aquesta pregunta fa referència als articles 10.5, 17.1 50.3 i 55.1 del Reglament 2024/1689 del Parlament i del Consell sobre intel·ligència artificial (AI Act), que versen sobre la governança de dades i el sistema de gestió de qualitat dels sistemes d'alt risc, així com les obligacions dels sistemes de reconeixement d'emocions, categorització biomètrica i de propòsit general de risc sistemàtic.

La pregunta també esta relacionada amb els articles 6 i 7 del Reglament 2016/679, de 27 d'abril 2016 [Reglament general de protecció de dades] que versen sobre la licitud del tractament de dades i el consentiment dels afectats, i l'article 21 de la Regulació 2022/868, de 30 de Maig del 2022 [Llei Europea sobre la Governança de dades], sobre els requisits específics per a protegir els drets i els interessos de les persones interessades i titulars de dades perquè respecta a les seves dades.

Així mateix, la pregunta es fonamenta en l'article 11 del Conveni marc del Consell d'Europa sobre intel·ligència artificial i drets humans, democràcia i estat de dret, que versa sobre la privacitat i protecció de dades personals; l'article 12 de la Declaració Universal dels Drets Humans, l'article 8 del Conveni Europeu de Drets Humans, i l'article 17 del Pacte Internacional de Drets Civils i Polítics, sobre el Dret humà a la privacitat; els articles 7 i 8 de la Carta de Drets Fonamentals de la UE sobre la privacitat i la protecció de dades personals; i l'article 7 del Conveni 108 per a la protecció de les persones físiques pel que fa al tractament de dades personals, que versa sobre la seguretat de les dades.

En aquest sentit, la qüestió guarda relació amb el principi *Privacy* del document A Human Rights-Based Approach To Data de l'ACNUDH i les recomanacions contingudes al *toolkit* d'ICO sobre riscos i protecció de dades en la IA.

Veure exemple

L'empresa *Meta* va ser multada el maig del 2023 per violar la Llei de Protecció de Dades incomplint amb la sol·licitud del consentiment de deure de protecció de les dades personals, tal com s'explica als següents articles: *This article is more than 4 months old Facebook owner Meta fined €1.2bn for mishandling user information, Meta Fined $1.3 Billion for Violating E.U. Data Privacy Rules* i *Meta slapped with record $1.3 billion EU fine over data privacy*.

SÍ	NO	NO APLICA

Els resultats

Mètriques (I): Resultats globals

Resultats globals

1 Transparència i explicabilitat
Preguntes contestades: 24 de 24 — 70% SÍ / 30% NO

2 Justícia i equitat
Preguntes contestades: 20 de 21 — 95% SÍ / 5% NO

3 Seguretat i no meleficència
Preguntes contestades: 20 de 20 — 100% SÍ / 0% NO

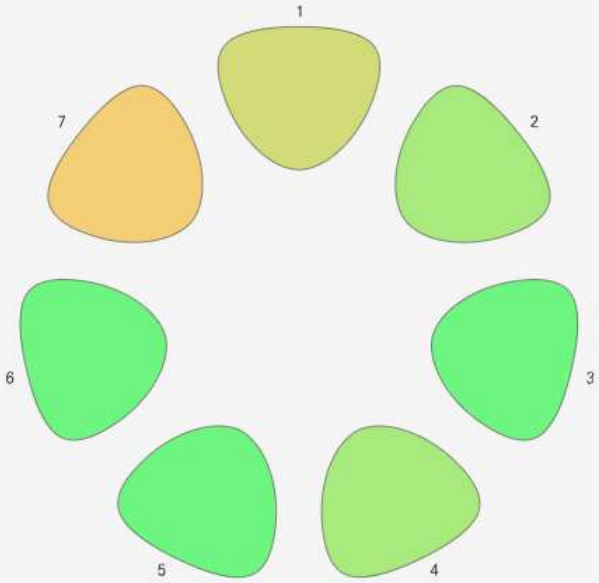
4 Responsabilitat i retiment de comptes
Preguntes contestades: 19 de 19 — 94% SÍ / 6% NO

5 Privacitat
Preguntes contestades: 19 de 19 — 100% SÍ / 0% NO

6 Autonomia
Preguntes contestades: 15 de 15 — 100% SÍ / 0% NO

7 Sostenibilitat
Preguntes contestades: 16 de 16 — 56% SÍ / 44% NO

0% 0-19% 20-39% 40-59% 60-79% 80-99%



Descarregar insígnia | Descarregar resultats

Mètriques (II): Matriu de riscos

Matriu de riscos

La següent matriu de riscos és una eina per obtenir una fotografia del nivell de risc segons la severitat (normativa o AI Act) i la probabilitat de portar a terme una acció al respecte. Els valors en cada casella representen les respostes negatives proporcionades segons la severitat i probabilitat esmentades. Aquesta informació també s'utilitza per als punts de millora.

	SEVERITAT SEGONS LA NORMATIVA O AI ACT		
PROBABILITAT DE PORTAR A TERME UNA ACCIÓ	Mínima	Limitada	Alta
Molt alta	1	0	4
Alta	2	0	4
Ni alta ni baixa	1	0	0
Baixa	1	1	2
Molt baixa	0	0	0

Mètriques (III): Punts de millora

Punts de millora

A continuació trobareu aquelles preguntes que heu contestat amb un "NO" i el seu grau de severitat. Els colors utilitzats en aquesta secció fan referència a la combinació donada entre els nivells de severitat (segons la normativa de la AI Act) i la probabilitat de que porteu a terme alguna acció aviat. Aquests apareixen en la mateixa taula anterior.

Molt baix | Baix | Baix/mitjà | Mitjà | Mitjà/alt | Alt

1 Transparència i explicabilitat

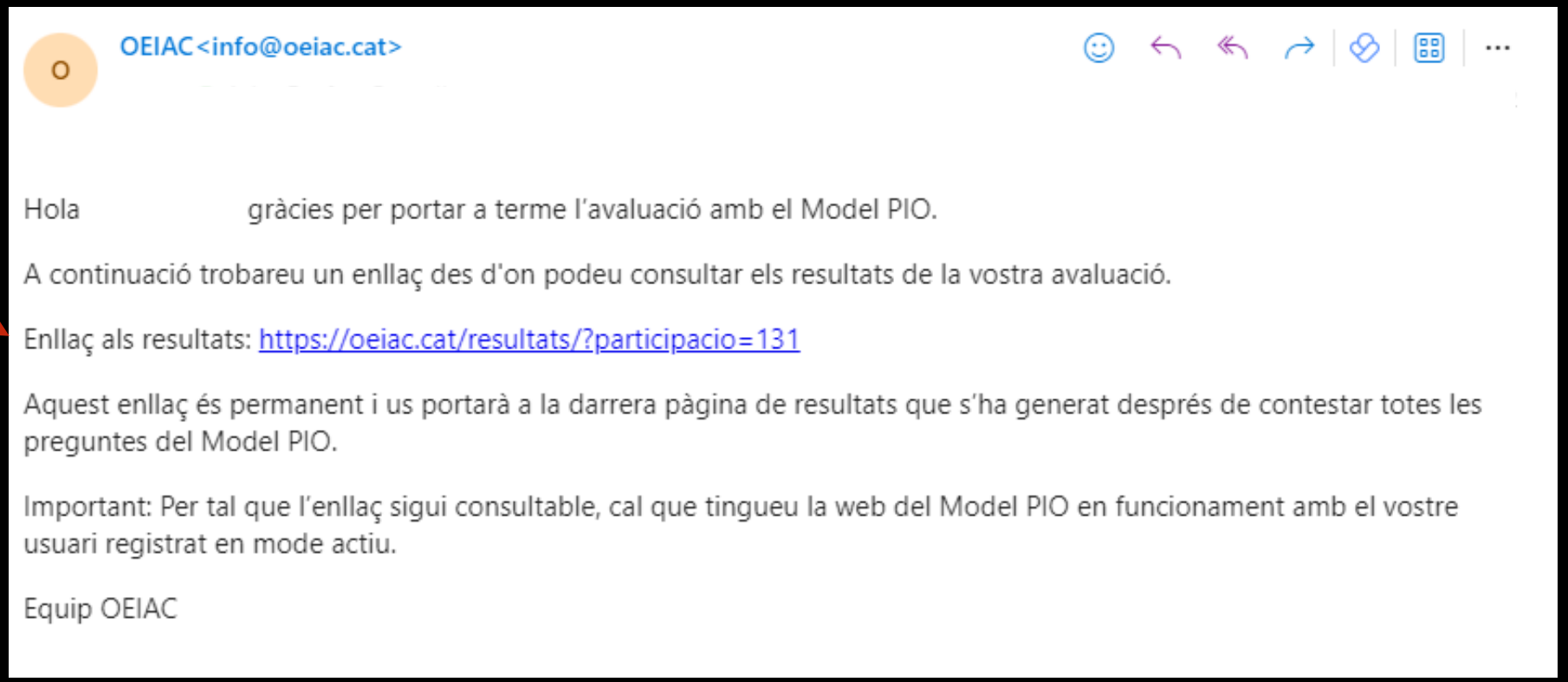
1.1.5 Heu facilitat o posat a disposició informació sobre la identitat i les dades de contacte de la persona proveïdora (o els seus representants)?	Aquesta pregunta es fonamenta en els articles 13.3, 22 i 54 del Reglament 2024/1689 del Parlament i del Consell sobre intel·ligència artificial (AI Act), que versen sobre la transparència i sobre les obligacions dels representants autoritzats del proveïdors de sistemes d'alt risc i els models de propòsit general. Així mateix, la pregunta es fonamenta en els articles 8 i 14 del Conveni marc del Consell d'Europa sobre intel·ligència artificial i drets humans, democràcia i estat de dret, que versen sobre la transparència i supervisió dels sistemes d'IA i els recursos per a les vulneracions dels drets humans derivades dels sistemes d'IA. Juntament amb el dret a la informació reconegut a l'article 19 del Pacte Internacional de Drets Civils i Polítics, l'article 11 de la Carta de Drets Fonamentals de la UE, l'article 10 del Conveni Europeu de Drets Humans i l'article 19 de la Declaració Universal dels Drets Humans.

Mètriques (III): Punts de millora

2 Justícia i equitat

2.1.1 Heu establert mesures d'inclusió i equitat com provar els resultats del vostre sistema d'IA per a diferents grups afectats?	Aquesta pregunta fa referència als articles 9, 10.3, 15.3, 27.1 i 95 del Reglament 2024/1689 del Parlament i del Consell sobre inteHligència artificial (AI Act), relatius a les mesures de gestió de riscos, la governança de dades, la resiliència i robustesa dels sistemes d'IA, l'avaluació d'impacte dels drets fonamentals per a sistemes d'IA d'alt risc i la promoció de codis de conducta d'aplicació voluntària per a persones o grups vulnerables. La pregunta també es relaciona amb els articles 10 i 16 del Conveni marc del Consell d'Europa sobre inteHligència artificial i drets humans, democràcia i estat de dret, que versen sobre la Igualtat i no discriminació i el marc de gestió de riscos i impactes. Juntament amb el dret humà la no discriminació, consagrat a l'article 14 del Conveni Europeu de Drets Humans, l'article 21 de la Carta de Drets Fonamentals de la UE, l'article 26 del Pacte Internacional de Drets Civils i Polítics i l'article 7 de la Declaració Universal dels Drets Humans. A més, la pregunta també està relacionada l'apartat *Specific safeguards to ensure non-discrimination when using* de l'informe *Getting The Future Right, Artificial Intelligence And Fundamental Rights* de la FRA. D'altra banda, la pregunta es fonamenta en l'apartat de *Fairness and non-discrimination* de les recomanacions sobre ètica en la InteHligència Artificial d'UNESCO, i està relacionada amb l'apartat *Fairness* del AI Ethics Assessment de GSMA.
2.1.2 Us heu assegurat que el vostre sistema d'IA no utilitza variables o "proxies" que poden ser injustament discriminatòries?	Aquesta pregunta fa referència als articles 9, 15.3 27.1, 55.1 i 95 del Reglament 2024/1689 del Parlament i del Consell sobre inteHligència artificial (AI Act), relatius a les mesures de gestió de riscos, la resiliència i robustesa dels sistemes d'IA, l'avaluació d'impacte dels drets fonamentals per a sistemes d'IA d'alt risc, les obligacions des proveïdors de sistemes de propòsit general de risc sistèmic i la promoció de codis de conducta d'aplicació voluntària per a persones o grups vulnerables. i la promoció de codis de conducta d'aplicació voluntària per a persones o grups vulnerables. A més, la pregunta es fonamenta en l'article 10 del Conveni marc del Consell d'Europa sobre inteHligència artificial i drets humans, democràcia i estat de dret, que versa sobre la Igualtat i no discriminació, així com el dret humà a la no discriminació, consagrat a l'article 14 del Conveni Europeu de Drets Humans, l'article 21 de la Carta de Drets Fonamentals de la UE, l'article 26 del Pacte Internacional de Drets Civils i Polítics i l'article 7 de la Declaració Universal dels Drets Humans. En aquest sentit, la pregunta també està relacionada amb l'apartat *Specific safeguards to ensure non-discrimination when using* de l'informe *Getting The Future Right, Artificial Intelligence And Fundamental Rights* de la FRA. Finalment, destaquem l'apartat de *Fairness and non-discrimination* de les recomanacions sobre ètica en la InteHligència Artificial d'UNESCO i l'apartat *Fairness* del *AI Ethics Assessment* de GSMA.
2.1.3 Heu avaluat si el vostre sistema d'IA s'adapta a un ampli espectre de preferències i habilitats individuals, incloent-hi usuaris que requereixen tecnologies d'assistència?	Aquesta pregunta fa referència a l'article 95 del Reglament 2024/1689 del Parlament i del Consell sobre inteHligència artificial (AI Act), relatiu a la promoció de codis de conducta d'aplicació voluntària en matèria d'accessibilitat. A més, la pregunta es fonamenta en l'article 10 del Conveni marc del Consell d'Europa sobre inteHligència artificial i drets humans, democràcia i estat de dret, que versa sobre la Igualtat i no discriminació, així com el dret humà a la no discriminació, consagrat a l'article 14 del Conveni Europeu de Drets Humans, l'article 21 de la Carta de Drets Fonamentals de la UE, i l'article 26 del Pacte Internacional de Drets Civils i Polítics. Destaquem també les recomanacions contingudes a l'apartat de *Diversity, Non-discrimination and Fairness: Accessibility and Universal Design* d'ALTAI, i està relacionada amb l'apartat *Fairness del AI Ethics Assessment de GSMA.*

Mètriques (IV)



O OEIAC<info@oeiac.cat>

Hola gràcies per portar a terme l'avaluació amb el Model PIO.

A continuació trobareu un enllaç des d'on podeu consultar els resultats de la vostra avaluació.

Enllaç als resultats: https://oeiac.cat/resultats/?participacio=131

Aquest enllaç és permanent i us portarà a la darrera pàgina de resultats que s'ha generat després de contestar totes les preguntes del Model PIO.

Important: Per tal que l'enllaç sigui consultable, cal que tingueu la web del Model PIO en funcionament amb el vostre usuari registrat en mode actiu.

Equip OEIAC

Altres recursos

Catàlegs: Exemples i Legal i recomanacions

A-B

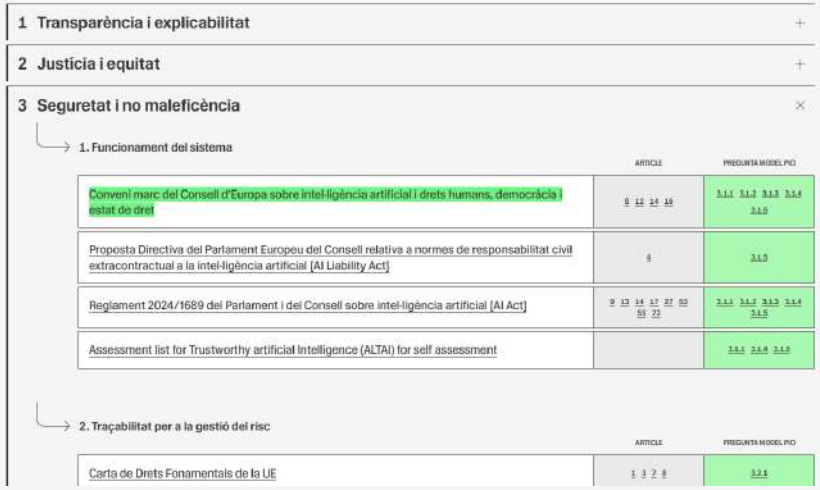
Exemples

Exemples reals i hipotètics a les preguntes del Model PIO.

100+

Legal i recomanacions

Normes jurídiques en les que s'han inspirat les preguntes del Model PIO.



Recomanacions de clàusules contractuals

Recomanacions

Oferim contingut recomanat per a la redacció de clàusules de contractes relatius a sistemes d'IA.

- Transparència
- Justícia
- Seguretat
- Responsabilitat
- Privacitat
- Autonomia
- Sostenibilitat
- Totes les recomanacions

Més recursos

La AI Act

Antecedents legislatius
Cronologia
Terminis d'aplicació
Altres dates rellevants
AI Liability Directive

Recursos

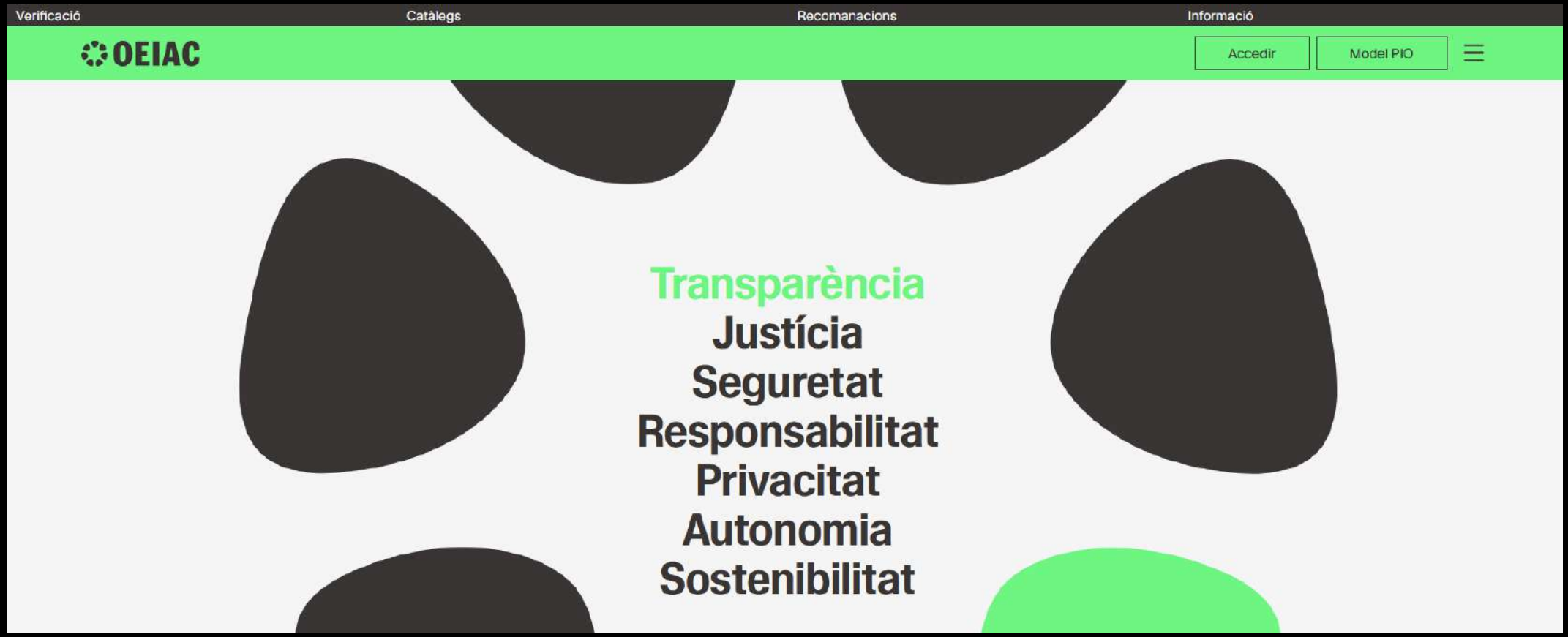
Avaluem
Acompanyem
Aconsellem
Formem

Avaluacions

Avaluació de conformitat
Avaluació d'impacte sobre drets fonamentals
Obligacions de transparència

Models sectorials

Els Models Sectorials són adaptacions del Model PIO que es desenvolupen amb diferents professionals i persones usuàries per donar resposta a les necessitats específiques d'àmbits o sectors concrets com la salut, l'educació, els serveis socials, etc. Aquest espai estarà dedicat a aquests Models Sectorials que estaran disponibles en els propers mesos en la seva versió beta.



www.oeiac.cat