# UNA APROXIMACIÓ A LA PRÀCTICA DE LA IA ÈTICA

Albert Sabater

Jornada organitzada per
Clúster TIC Turisme – ACCIO / Direcció General de Turisme
19/4/2021 | 10:00h – 11:30h

# 3 àrees diferenciades i interrelacionades

Focus en la previsió i regulació
*Foresight AI and regulation*


Focus en el desenvolupament
*Design principles for ethical AI*

Focus en la implantació i monitoratge
*Understanding AI social impacts in diverse social settings*

# Sumari

1. A què ens referim per ètica de la IA?
2. Per què l'emergència de la IA ètica?
3. Quins són els principals riscos de la IA?
4. La IA en el sector turístic
5. Com es pot aplicar una IA ètica?

# A què ens referim per ètica de la IA? (I)

- L'ètica de la IA fa referència a les **dues preocupacions.**

1) Una sobre el comportament moral de les persones en el disseny, fabricació i utilització de sistemes d'IA (ethics of technology).

2) Una altra sobre el comportament dels sistemes d'IA (machine ethics).

- La ètica de la IA vol **garantir** que tant el comportament dels tecnòlegs com el comportament dels sistemes d'IA no perjudiquin a les persones (tampoc al seu entorn i altres éssers vius i els seus hàbitats).

# A què ens referim per ètica de la IA? (II)

- En l'ètica de la IA la qüestió central no és si la IA pot fer una o altra cosa, **la qüestió és si l'haurien de fer i com**.

- Per tant, es pot dir que l'ètica de la IA busca resoldre qüestions de **moralitat humana** i d'**acceptació social**, tenint en compte el que és susceptible de fer el bé i el mal, i prenent en consideració principis morals com la responsabilitat, la justícia, la confiança, la transparència, la inclusivitat, la sostenibilitat, entre d'altres.

- Però com qualsevol tecnologia, l'adopció i aplicació difereix segons el **context**.

# Per què l'emergència de la IA ètica? (I)

- Cal situar-la en el progrés de la Quarta Revolució Industrial, amb IA, dades massives, robòtica i Internet de les Coses (IoT).

- Quant a la IA, les preocupacions estan centrades en 3 tipus de progrés (Narayanan, 2019):

1. **Progrés tecnològic (genuí i ràpid) en temes de percepció:**

- Identificació de continguts.
- Reconeixement facial*.
- Diagnosi mèdica a través d'escanejats.
- Reconeixement de la parla (*speech to text*).
- *Deepfakes**.

***Preocupació ètica i social per raons de precisió**

# Per què l'emergència de la IA ètica? (II)

2. **Progrés tecnològic (lluny de ser perfecte però millorant) en automatització de criteris o resolució de problemes:**

- Detecció de *spam*.
- Detecció de material amb drets d'autor/a.
- Automatització d'avaluacions.
- Detecció de discursos d'odi.
- Sistemes de recomanació de continguts.
- Xatbots i assistents virtuals.

**Preocupació ètica i social per errors inevitables**

→ S'utilitza l'heurística, un mecanisme cognitiu que ajuda a trobar respostes correctes però imperfectes a preguntes complexes.

# Per què l'emergència de la IA ètica? (III)

3. **Progrés tecnològic (fonamentalment dubtós) en la predicció de fenòmens de naturalesa social:**

- Predicció de la reincidència criminal.
- Predicció del rendiment laboral.
- Policia predictiva.
- Predicció del risc terrorista.
- Predicció de persones en risc.

**Preocupació ètica i social amplificada per la inexactitud**

→ L'error de voler predir comportaments (individuals o de grup) a partir d'una relació/associació no causal (X~Y no és X→Y).

# Quins són els principals riscos de la IA?

- La IA genera un ampli debat ètic tan en termes de desenvolupament ("Choice Development") com d'implantació ("Impact and Consequences").

- En aquest context, els principals riscos que se'n deriven són:

1. L'abús de dades massives (**Big Data**) i el biaix algorítmic (**AI bias**).

2. Les caixes negres de la IA (**Black box AI**).

# Big Data & AI bias (I)

- Molts sistemes d'IA, com ara els d'aprenentatge automàtic supervisat (ML o Machine Learning), **depenen** de grans quantitats de dades per funcionar bé.
- S'assumeix (però no per molt temps!) que com més dades millor.
- **"Garbage In Garbage Out" (GIGO)**: Si les dades no són bones, es a dir, esbiaixades o poc representatives, els algoritmes no aporten bones solucions ans el contrari.
- En gran part el problema de les solucions esbiaixades deriva de les males dades o **Big Bad Data**.

# Big data & AI bias (II)

- És per això que, de moment, s'han plantejat tres solucions parcials:

1. **De "Big Data" a "Good Data"**, que vol dir utilitzar dades massives (o no) que siguin el més representatives possible.

2. **Promoure una IA diversa i inclusiva**, incloent-hi comunitats poc representades per visibilitzar actors i dades i evitar la discriminació.

3. **Comprendre, mesurar i mitigar els biaixos**, incloent-hi o no solucions d'IA.

# Black box AI (I)

- Moltes tècniques d'aprenentatge automàtic (Machine Learning) o profund (Deep Learning) consisteixen en utilitzar algoritmes i milions de dades en processos autodirigits que es consideren **caixes negres**.

- El problema és que moltes vegades és **difícil entendre o remuntar-se al procés** pels quals arriben a determinades solucions o prediccions, raó per la qual s'han guanyat el nom de caixes negres.

- Principalment es tracta d'un **problema tècnic** ja que l'ús de ML o DL i encara més les xarxes neuronals profundes (DNN) poden ser tan difícils d'entendre com el cervell humà.

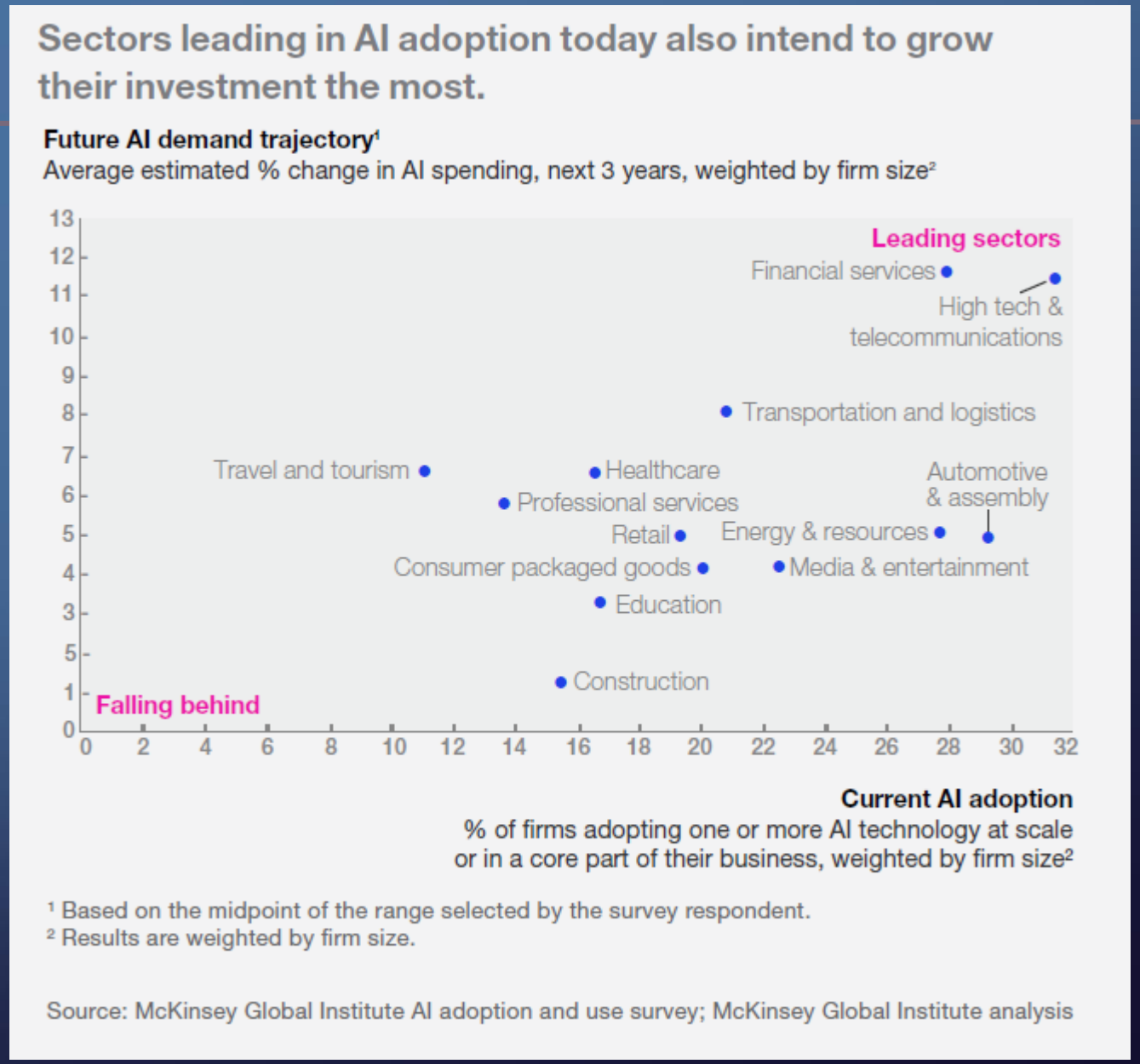
# Black box AI (II)

- 4 preguntes clau a recordar:

1. El sistema d'IA funciona tal i com es pretén o no? [SEGURETAT TÈCNICA]
2. Si funciona, podem comprendre el seu funcionament? [TRANSPARÈNCIA]
3. Qui és responsable si deixa de funcionar correctament? [RESPONSABILITAT]
4. Quin és l'impacte si deixa de funcionar correctament? [JUSTÍCIA]

# IA en el sector turístic

- Segons l'estudi del McKinsey Global Institute (2018), el nivell d'adopció global en el sector turístic ha estat molt baix fins ara.
- No obstant, el sector està augmentant molt la despesa relacionada amb la IA i es preveu que passi de la posició 13 a la 4 de les indústries analitzades.



# Els tres principals usos de la IA en el sector turístic

1. **En aplicacions d'atenció i informació** al client.
2. **En la fase de cerca i reserva** del viatge del client.
3. **En la fase d'experiència** mitjançant robots.

- Les eines principals que s'utilitzen són els xatbots i assistents virtuals, els sistemes de recomanació i personalització, els sistemes de predicció i els robots de servei.

# L'exemple del xatbot

- En desenvolupar-lo, dos aspectes claus són la **privadesa** i la **transparencia**.

- Això ajuda a respondre a preguntes com ara:

-On s'emmagatzemen les dades de la transcripció del xat?
-Es poden estudiar les converses amb un xatbot per millorar l'experiència de l'usuari?
-Quant de temps s'ha de conservar la transcripció del xat?
-Si es planteja una queixa mitjançant el xatbot, qui ho supervisarà?

- També cal comptar amb la **responsabilitat** i la **justicia**. Per exemple, tindrà el xatbot una identitat masculina, femenina o cap? El mateix és aplicable a l'origen.

# Com es pot aplicar una IA ètica?

- Com podem operacionalitzar diferents valors més enllà de la seva comprensió?

| | |
|---|---|
| **Responsabilitat** | **Transparència** |
| **Justícia** | **Seguretat** |

- **Els sistemes d'IA han de ser dissenyats i aplicats per persones i aquestes han de ser les úniques responsables.**
- **Els sistemes d'IA han de tractar a tothom d'una manera justa i sense discriminar a cap persona.**
- **Els sistemes d'IA han de ser transparents i que es puguin explicar de manera general i detallada.**
- **Els sistemes d'IA han de ser segurs i garantir la privacitat de les persones.**

# Combinació de tres eines

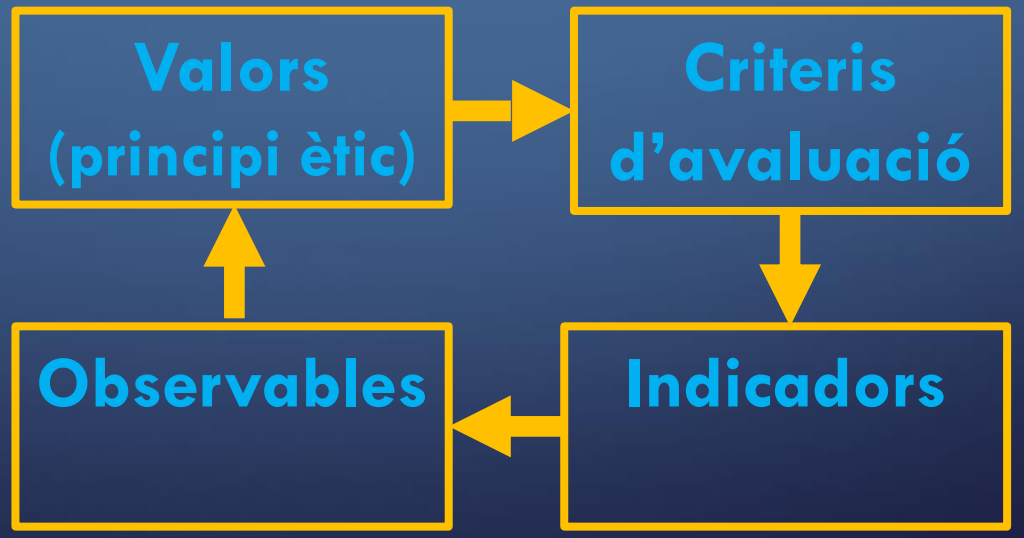
**1 Model VCIO**
Valors, Criteris, Indicadors, Observables

**2 Classificació**
de propietats del Model VCIO

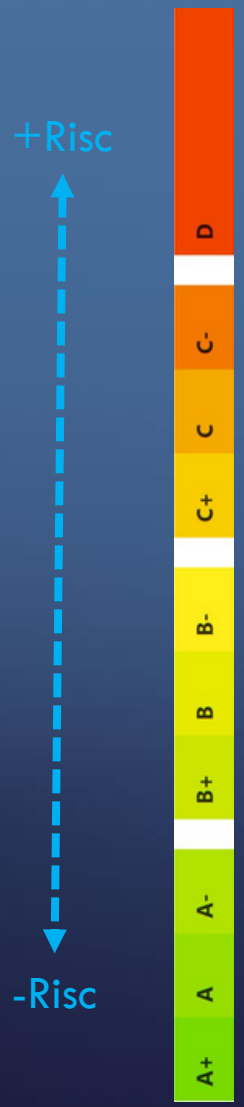
**3 Etiquetatge**
de IA ètica basada en el model VCIO

# Una aproximació: El model VCIO

- Algunes organitzacions com el AI Ethics Impact Group (AIEIG) utilitzen l'anomenat **model VCIO (Valors, Criteris, Indicadors, Observables)** per tal de fer practicables, comparables i mesurables els principis ètics.



# Un exemple VCIO: Transparència (I)




-No

-Si però

-Si

**Transparència (Valor)**

**Criteris d'avaluació**

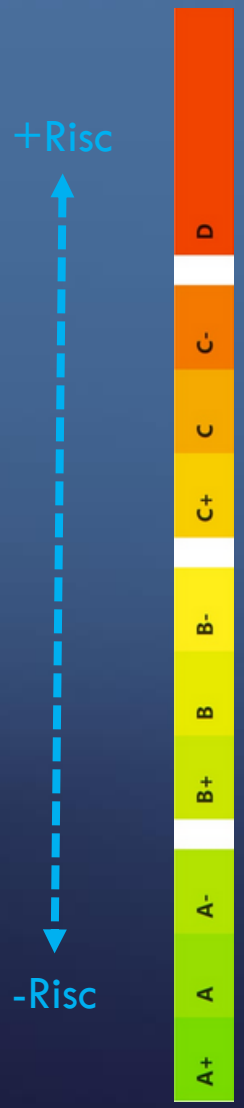
-Origen/destinació dels conjunts de dades

**Indicadors**

**Observables**

-L'origen de les dades està documentat?
-Per a cada propòsit, sabem quines dades s'utilitzen?
-Es monitoritza la destinació i ús de les dades?

# Un exemple VCIO: Transparència (II)




-No

-Si però

-Si

**Transparència (Valor)**

**Criteris d'avaluació**

-Propietats dels algoritmes o models

**Indicadors**

**Observables**

-El model en qüestió s'ha provat i utilitzat abans?
-És possible inspeccionar-lo per possibles debilitats?
-Tenint en compte l'eficiència i la precisió, s'ha utilitzat el model més senzill?

# Una aproximació: El model VCIO

- Els resultats en els observables determina la posició de qualificació del sistema IA.
- Diferents etiquetatges possibles (estàndard o ad hoc).




Source: AI Ethics Impact Group (AIEIG)

# Una aproximació: El model VCIO

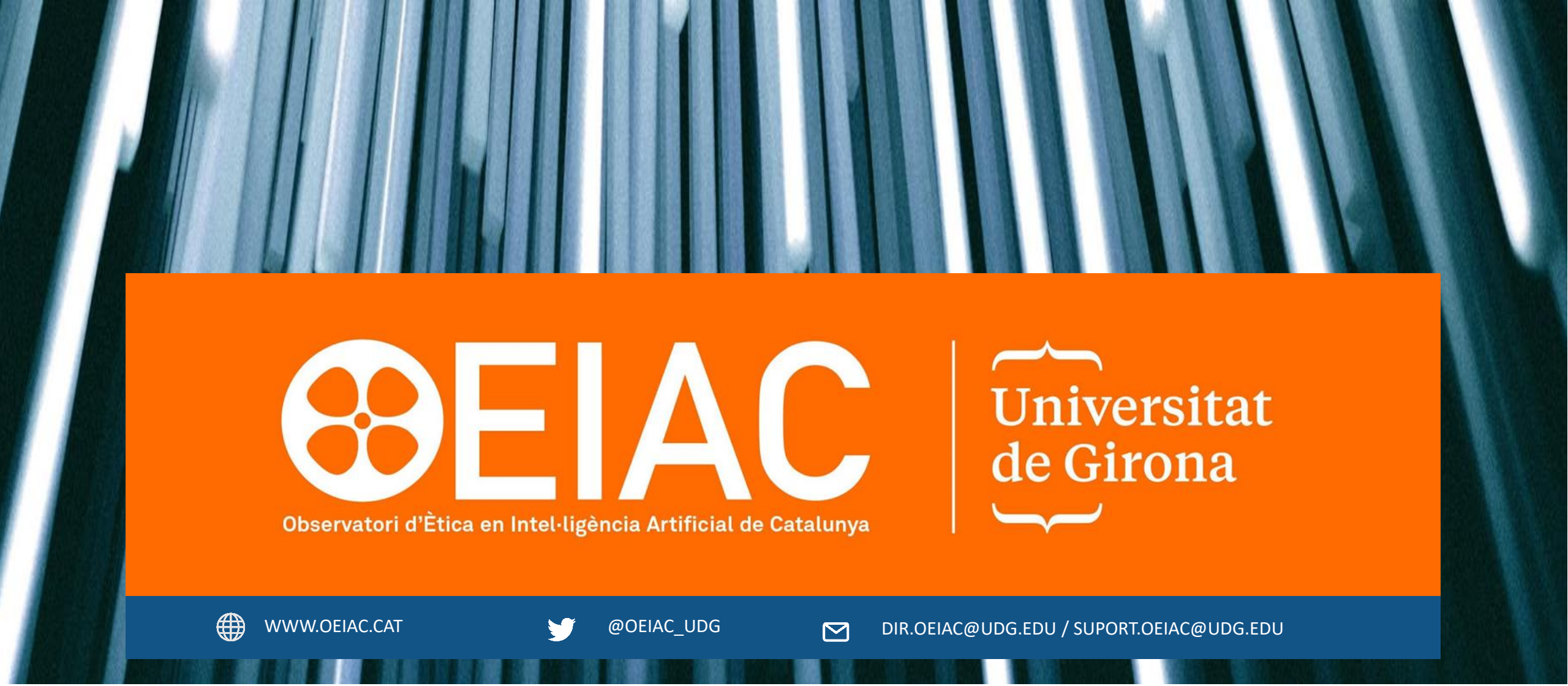
- Els requisits mínims que els sistema d'IA ha de complir depenen del sector.
- Per exemple, A- podria ser suficient per turisme però no pas per salut.




Source: AI Ethics Impact Group (AIEIG)

# No tot és observable i quantificable

- El model VCIO és pot millorar ja que **no tot és observable i quantificable**.
- La mètrica no és important per si mateixa i cal tractar-la com una aproximació.
- Hi haurà desenvolupaments de la IA que tenen riscos intangibles que requereixen una revisió **més enllà de les mètriques**.
- Combinació de mètriques amb avaluacions qualitatives, incloent-hi la inclusió persones o grups potencialment afectats per la implementació d'un sistema d'IA.



OEIAC

Observatori d'Ètica en Intel·ligència Artificial de Catalunya

Universitat de Girona

WWW.OEIAC.CAT

@OEIAC_UDG

DIR.OEIAC@UDG.EDU / SUPORT.OEIAC@UDG.EDU

Universitat de Girona

Generalitat de Catalunya
Departament de Polítiques Digitals i Administració Pública

i2cat