DE LA CONCEPCIÓ A L'ACCIÓ

UNA APROXIMACIÓ A LA PRÀCTICA DE LA IA ÈTICA

Albert Sabater

CIDAI-Masterclass Everis – AI Ethics & Society
5/3/2021 | 13:30h – 15h

Preàmbul

"Any sufficiently advanced technology is indistinguishable from magic"
Arthur C. Clark

"With great power comes great responsability"
Peter Parker principle

- Gran augment de la mida dels conjunts de dades.
- Enorme augment de la potència de computació.
- Gran millora en els algoritmes de ML.

Sumari

1. Els principals riscos de la IA
2. 4 preguntes clau a recordar
3. El problema de la concepció a l'acció
4. Una aproximació: El model VCIO i una matriu d'aplicació
5. Alguns aspectes a considerar i algunes solucions

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**).

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]

El problema de la concepció a l'acció (I)

- Mentre les idees generals d'**adopció** de principis ètics en la IA comença ha ser **generalitzada**, l'**aplicació** de principis ètics està sent molt **més difícil** d'implementar degut a la **comprensió variable** i a la **manca d'estàndards** per la seva avaluació.

Transparència Responsabilitat
Explicabilitat Confiança Justícia
Solidaritat Equitat Privadesa
Benestar Dignitat Seguretat
Autonomia Inclusivitat Diversitat
Sostenibilitat Prevenció Fiabilitat

El problema de la concepció a l'acció (II)

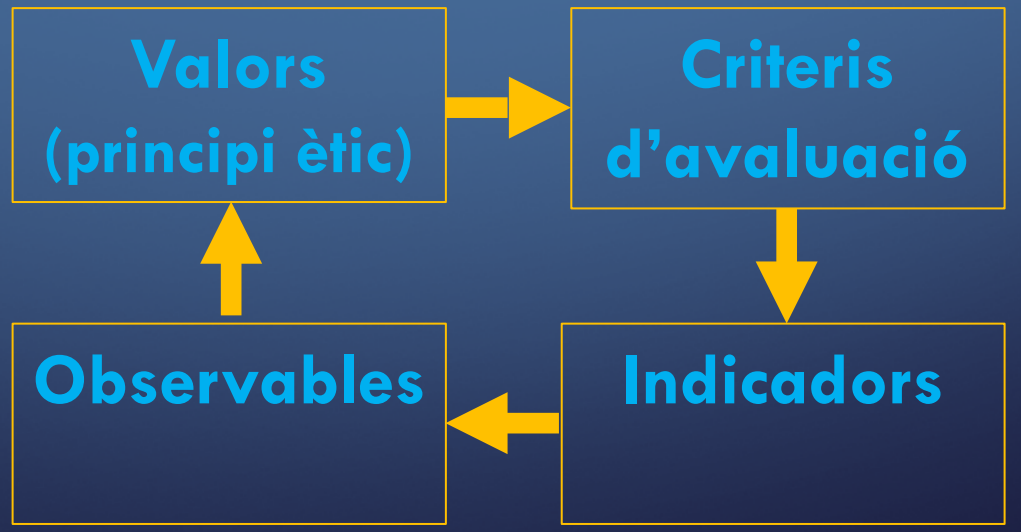
- Com podem operacionalitzar-los més enllà de la seva comprensió?

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.

Una aproximació: El model VCIO (I)

- 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)

Transparència (Valor)

Criteris d'avaluació

-Divulgació de l'origen dels conjunts de dades

Indicadors

Observables

-Si

-Si però

-No

-L'origen de les dades està documentat?
-Per a cada propòsit, sabem quines dades s'utilitzen?
-Estan documentades i divulgades?

Un exemple VCIO: Transparència (II)

-Si

-Si però

-No

Transparència (Valor)

Criteris d'avaluació

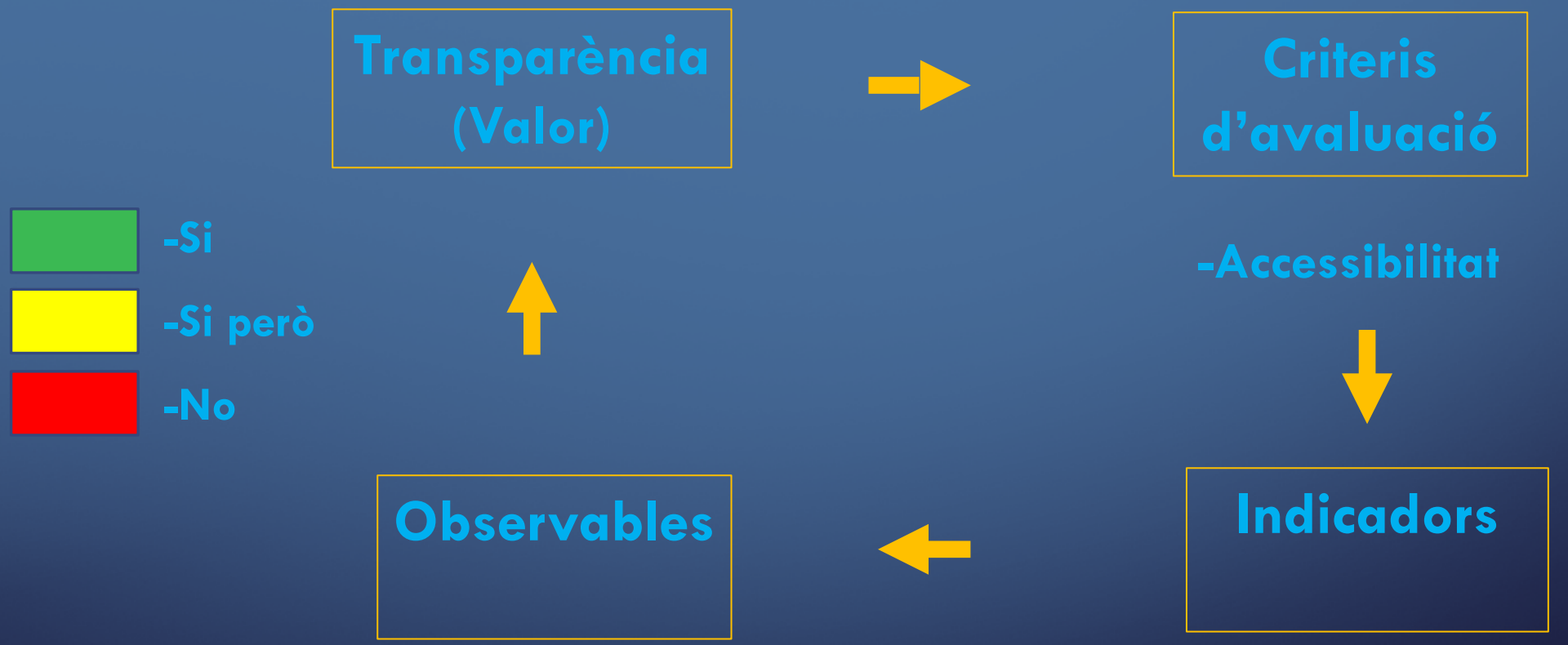
-Divulgació de propietats de l'algorisme o model

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 i intel·ligible?

Un exemple VCIO: Transparència (III)



-El principi de funcionament és comprensible i interpretable?
-S'ha creat una autoritat mediadora per resoldre i regular conflictes de transparència?

Una aproximació: El model VCIO (II)

- Els resultats en els observables (Si, Si però, i No) determina la posició de qualificació (A = millor valoració, i G = pitjor valoració) dels sistema IA.

Source: AI Ethics Impact Group (AIEIG)

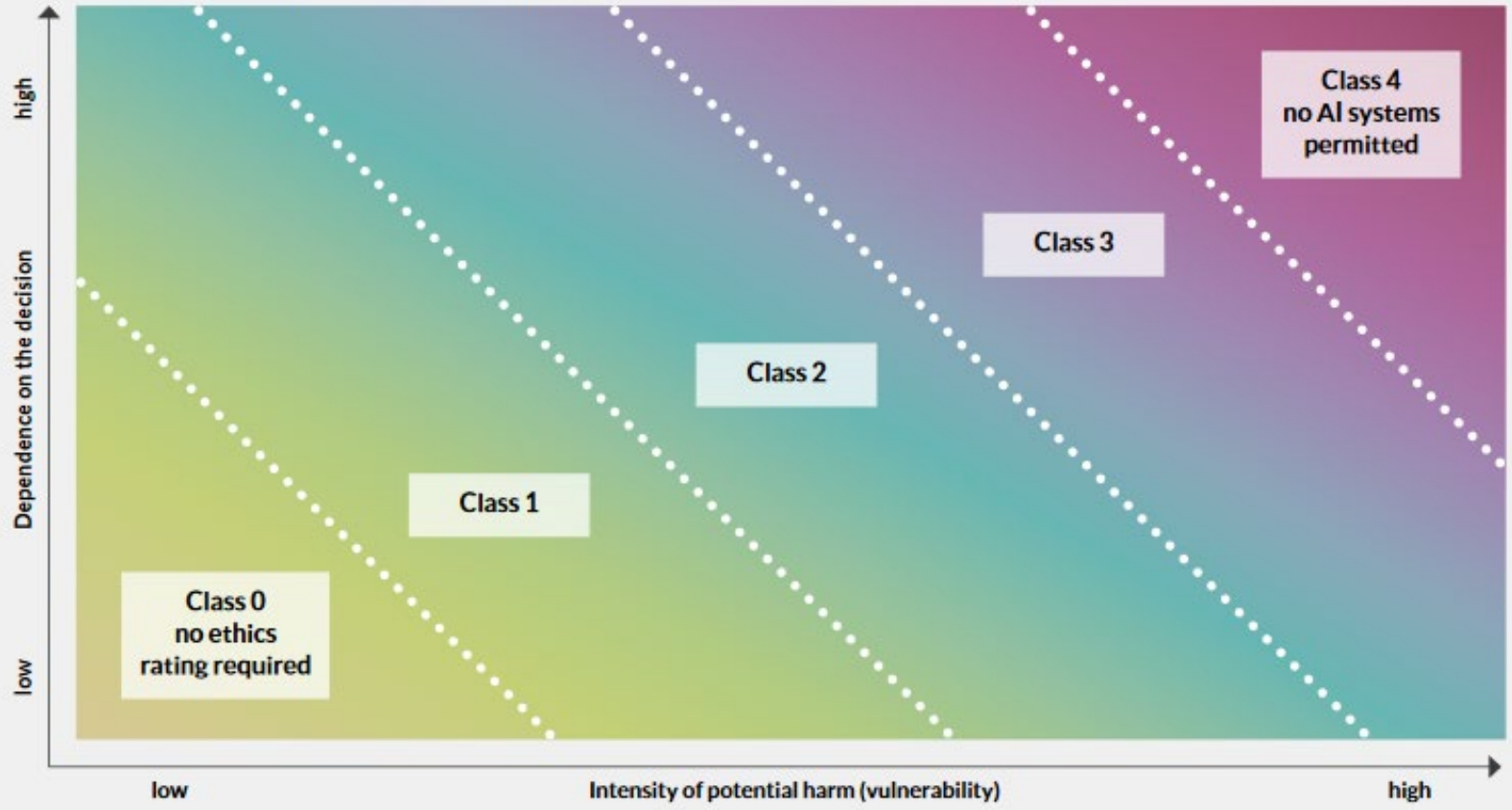
Una aproximació: El model VCIO (III)

- Per determinar els requisits mínims que els sistema d'IA ha de complir en un context d'aplicació específic. Per exemple, C en transparència podria ser suficient en l'àmbit de la mobilitat però no pas en el de la salut.

Source: AI Ethics Impact Group (AIEIG)

Una matriu d'aplicació

FIGURE 7 **Risk matrix with 5 classes of application areas with risk potential ranging from 'no ethics rating required' in class 0 to the prohibition of AI systems in class 4**



Source: Krafft and Zweig 2019

AIEI Group

Una aproximació: El model VCIO (IV)

- Seguint aquest model VCIO, es proposa la creació d'una **etiqueta de valors per als sistemes de IA**, similar a l'etiqueta d'eficiència energètica de la Unió Europea per a electrodomèstics.

- Aquesta etiqueta podria **millorar la comparabilitat i avaluació ràpida** sobre si un sistema determinat compleix els requisits ètics necessaris per a una aplicació determinada.

- Mitjançant aquests mecanismes, l'enfocament podria **incentivar el desenvolupament ètic de la IA** més enllà dels requisits actualment recollits a la llei.

No tot és observable i quantificable

- El model VCIO és pot millorar ja que **no tot és observable i quantificable**.
- Hi haurà desenvolupaments de la IA que tenen riscos intangibles que requereixen una revisió acurada i **crítica més enllà de les mètriques**.
- També es requereixen **mecanismes de rendició de comptes, documentació i models d'inventaris** per desplegar la IA de manera ètica.

D'altres aspectes a considerar

- Qualsevol mètrica és poc útil fora del seu context.
- La mètrica no és important per si mateixa i cal tractar-la com una aproximació.

"When a measure becomes a target, it ceases to be a good measure" (Goodhart 2015 and Strathern 1997).

- Del contrari els intents de manipular-les esdevenen habituals (**gaming**).
- La mètrica ha d'anar més enllà del curt termini (**short-term**).

Algunes solucions

- Utilitzar mètriques tenint en compte el context d'aplicació (**context-based**)
- Utilitzar un ventall de mètriques per una imatge més completa (**multiple proxies**)
- Combinació de mètriques amb avaluacions qualitatives (**game avoidance**)
- Implicar diferents grups d'interès, i els afectats (**expansion of ethical circle**)
- Utilitzar mètriques consistents al llarg del temps (**comparability**)

WWW.OEIAC.CAT @OEIAC_UDG SUPORT.OEIAC@UDG.EDU

CIDAI-Masterclass Everis – AI Ethics & Society

5/3/2021 | 13:30h – 15h