EIAC
Observatori d'Ètica en Intel·ligència Artificial de Catalunya

El model PIO

(Principis, Indicadors i Observables):

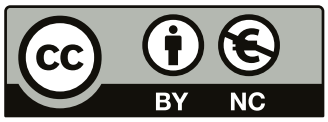
Una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'intel·ligència artificial

Universitat de Girona

Generalitat de Catalunya
Departament de la Vicepresidència
i de Polítiques Digitals i Territori

i2cat

El Model PIO (Principis, Indicadors i Observables):
Una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'intel·ligència artificial



Juny de 2022

Edició

Observatori d'Ètica en Intel·ligència Artificial de Catalunya (OEIAC)

Carrer de la Universitat, 10 – Edifici Econòmiques

Universitat de Girona

17003 Girona

Redacció

Albert Sabater (OEIAC) i Alicia de Manuel (OEIAC).

El model PIO (Principis, Indicadors i Observables):

Una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'intel·ligència artificial

2022

OBSERVATORI D'ÈTICA EN INTEL·LIGÈNCIA ARTIFICIAL DE CATALUNYA

Índex.

1. Introducció i objectius **7**

2. El principis ètics del model PIO **10**

2.1. Transparència i explicabilitat **12**

2.2. Justícia i equitat **14**

2.3. Seguretat i no maleficència **16**

2.4. Responsabilitat i rendició de comptes **17**

2.5. Privacitat **18**

2.6. Autonomia **19**

2.7. Sostenibilitat **20**

3. Definició, eines i model PIO (Principis, Indicadors i Observables) **22**

3.1. Toolkits i checklists ètics **26**

3.2. Certificacions ètiques **33**

3.3. La proposta i metodologia del model d'autoavaluació PIO **39**

3.4. Continguts del qüestionari d'autoavaluació del model PIO **45**

Transparència i explicabilitat **48**

1. Anàlisi preliminar de l'ús del sistema
2. Explicabilitat
3. Traçabilitat i representativitat
4. Protocols de comunicació

Justícia i equitat **51**

1. Inclusió i equitat
2. Identificació de biaixos

Seguretat i no maleficència **53**

1. Funcionament del sistema
2. Traçabilitat per a la gestió del risc
3. Mecanismes de seguretat
4. Protecció enfront d'atacs

Responsabilitat i rendició de comptes **56**

1. Control
2. Retiment de comptes
3. Auditabilitat
4. Figures de protecció

Privacitat **58**

1. Protecció de la privacitat
2. Governança de dades

Autonomia **60**

1. Impacte en les capacitats dels usuaris
2. Percepció del sistema

Sostenibilitat **62**

1. Desenvolupament sostenible
2. Impacte mediambiental

Agraïments **78**

Bibliografia **79**

1 Introducció i objectius

Les consideracions ètiques al voltant de la intel·ligència artificial (IA) es remunten al principi de la informàtica digital (Wiener, 1954) i, és clar, les preguntes sobre ètica associades a la tecnologia tenen dècades (o segles) d'història[1]. Però el fet que hi hagi hi ha una sèrie de desenvolupaments recents en la IA que tenen una rellevància social més àmplia, ha fet que el debat i avaluació dels usos ètics de la IA vagi molt més enllà d'un cercle reduït de persones interessades en la tecnologia. En aquest sentit, la creixent importància dels sistemes d'IA per a nombrosos aspectes de la nostra vida quotidiana ha fet que es demani una major inclusió de consideracions ètiques i socials. I, és clar, per fer-ho hem de tenir clar quins són els aspectes que hem de tenir en compte i aplicar-los d'una manera senzilla i pràctica per tal d'exercir un continu procés de rendició de comptes dels usos de la IA (Mittelstad, 2019). No obstant, són molt poques les orientacions sobre com s'integren les preocupacions ètiques i socials en ecosistemes d'innovació malgrat la seva importància per avançar i obrir el camí per a una IA realment de confiança[2].

Juntament amb la proliferació i el convenciment que existeixen massa guies de principis ètics (Jobin et al., 2019), un altre problema comú és que les directrius sobre les consideracions ètiques que estan disponibles són sovint massa complexes i difícils d'entendre ja que estan escrites per a usuaris acadèmics i/o tècnics. Això no tan sols dificulta la seva comprensió per a diferents interessades com les empreses, l'administració, els organismes professionals i organitzacions de la societat civil, sinó -encara més greu- retarda la seva adopció i aplicació com un element de valor d'innovació responsable (Morley et al., 2019). A la discussió sobre els usos ètics de la IA per explorar com haurien de ser els ecosistemes responsables d'innovació d'IA, se li suma la poca o nul·la representació de persones usuàries, destinatàries o afectades pels propis sistemes d'IA. En conseqüència, s'ha convertit en una prioritat obrir el focus de la IA i traduir principis ètics en indicadors i observables de fàcil interpretació que proporcionin mesures quantitatives i qualitatives sobre l'ús ètic de dades i sistemes d'IA.

[1] Un dels primers temes és la tesi que la tecnologia aprèn o imita la natura, tal i com apareix en Les Lleis de Plató (vol. X), en un diàleg ambientat a l'illa grega de Creta al segle IV a. C. Mentre que el coetani Aristòtil es refereix a aquest tema en el llibre La Física, on sosté (repetint a Demòcrit) que: "generalment la technè en alguns casos completa allò que la natura no pot acabar, i en altres imita la natura".

[2] European Commission. (2019, 8 d'abril) "Ethics Guidelines for Trustworthy AI – Independent High-Level Expert Group on Artificial Intelligence" https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (consultada l'11 de març de 2022)

Des de 2017, quan el Canadà es va convertir en el primer país a adoptar una estratègia nacional d'intel·ligència artificial, almenys 60 països han adoptat polítiques (declaracions, marcs, directrius de la indústria o principis) centrades en la IA, per a treballar juntament amb el sector tecnològic, l'acadèmia i la societat civil per a un disseny i aplicació de la IA responsable, de confiança i ètic. En aquest sentit, a Catalunya es va posar en marxa al febrer de 2020 *l'Estratègia d'Intel·ligència Artificial de Catalunya* (CATALONIA.AI, 2020), per a enfortir l'ecosistema català en IA tenint molt present un eix sobre "Ètica i Societat". Per desenvolupar aquest eix, l'Observatori d'Ètica en Intel·ligència Artificial de Catalunya (OEIAC) ha iniciat una sèrie d'actuacions, entre les quals es troba la publicació d'informes com el recent sobre *Intel·ligència artificial, ètica i societat: Una mirada i discussió a través de la literatura especialitzada i d'opinions expertes* o el present informe sobre la importància no sols de prendre consideracions ètiques i consciència de l'impacte social de la IA a nivell teòric, sinó també sobre la necessitat d'avançar cap a la seva aplicabilitat.

En aquest sentit, el present document representa una guia d'implementació i autoavaluació per a organitzacions amb dos objectius clars:

- Sensibilitzar als diferents agents de la quàdruple hèlix, és a dir, la presència dels quatre pilars clau en qualsevol procés innovador: universitats i centres de recerca, administració pública, teixit empresarial, i ciutadania, que utilitzen dades i sistemes d'intel·ligència artificial, sobre la importància d'adoptar principis ètics fonamentals per minimitzar riscos coneguts i desconeguts i maximitzar oportunitats.

- Identificar accions adequades o inadequades a través d'una proposta d'autoavaluació basada en principis, indicadors i observables per valoritzar, recomanar i avançar en l'ús ètic de dades i sistemes d'intel·ligència artificial.

Aquests objectius parteixen d'un fet incontestable, que és la proliferació de productes i serveis que utilitzen sistemes d'IA. Davant d'això, el públic ha d'assegurar-se que els sistemes d'IA són, entre altres, justos, explicables i segurs, i que les organitzacions que els implementen són transparents i responsables. Així doncs, per fomentar la confiança del públic en les tecnologies d'IA cal un suport creixent a les estratègies d'autoavaluació sistemàtica per ajudar a aconseguir resultats més segurs, més fiables i més justos; per reduir el risc d'impacte negatiu en els afectats per les aplicacions d'IA; i per posar en pràctica generalitzada els estàndards ètics més alts per part d'organitzacions a l'hora de dissenyar, desenvolupar i implementar sistemes d'IA.

El model PIO és un pas més per avançar en aquesta direcció ja que proporciona una integració de la teoria a la pràctica de principis ètics fonamentals de la IA, i es construeix al voltant de la senzillesa i d'una pregunta clau i efectiva que qualsevol persona que desenvolupa, gestiona o dirigeix un projecte on hi ha dades i sistemes d'IA pot entendre

fàcilment: Ho hem fet? En aquest sentit, partint d'una perspectiva de responsabilitat social organitzativa, es proposa la utilització d'un model que es pot utilitzar en qualsevol fase del cicle de vida de les dades i sistemes d'IA. Això significa que és aplicable en el disseny i modelització, incloent-hi planificació, recollida de dades i construcció de models; en el desenvolupament i validació, incloent-hi formació de models i proves; i en el desplegament, seguiment i perfeccionament, incloent-hi en la resolució de problemes.

Volem deixar clar des d'un bon principi, que el model PIO no proporciona un instrument legal d'autoavaluació, i que les organitzacions tenen la flexibilitat per triar com s'acosten al model. No obstant, és clau compartir els mateixos objectius finals de millora a través de la identificació dels reptes per implementar els principis ètics en les dades i sistemes d'IA. És probable que algunes organitzacions ja tinguin marcs d'ètica més o menys formals i, per tant, el model PIO de l'OEIAC pot servir per avaluar les seves pràctiques existents i per identificar similituds i llacunes. També hi ha la possibilitat que algunes organitzacions vulguin avaluar alguns principis ètics tenint en compte aplicacions específiques, com la transparència i explicabilitat, la responsabilitat i retiment de comptes o la privacitat, entre altres, per una qüestió d'interès en una àrea concreta d'actuació de l'organització. Aquest darrer aspecte és clau no tan sols en termes de flexibilitat o especificitat de cadascuna de les organitzacions i aplicacions, sinó també per la millora del model PIO, que no s'entén com un tot i tampoc funciona d'una manera aïllada. En aquest context, és important assenyalar que el model PIO és dinàmic, ja que al marge de proporcionar informació i comentaris sobre àrees que necessiten una major orientació, també pretén integrar coneixements i compartir aprenentatges i consells de les experiències que s'aniran acumulant entre diferents organitzacions a través de casos pràctics. I a qui està dirigit aquest treball en general i el model PIO en particular? La resposta és senzilla: a totes les organitzacions públiques o privades que tinguin un interès en el disseny, desenvolupament o desplegament de tecnologies d'IA, i a totes les persones que estiguin utilitzant, comprant o siguin destinatàries d'aquestes tecnologies i plantegin preguntes com: Quins són els riscos i beneficis d'aquestes tecnologies? Qui es veu afectat per elles i com? De quina manera es pot millorar el benestar de les persones amb aquestes tecnologies? Creiem que aquestes i altres preguntes són prou rellevants per interpel·lar a tothom.

A continuació oferim una panoràmica i justificació dels principis ètics escollits del model PIO. Després aportem informació sobre algunes propostes i eines existents com els toolkits, checklists i certificacions ètiques per tal que quedi constància de la revisió anterior al desenvolupament del model PIO i, és clar, per confirmar que aquest no s'ha confeccionat de manera aïllada sinó pensant amb nombroses iniciatives existents i emergents. La secció més gran d'aquest treball fa referència a la presentació i descripció del *Model PIO (Principis, Indicadors, Observables): Una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'intel·ligència artificial*, a través d'un mòdul d'implementació en un entorn web [www.oeiac.cat].

2022

2 El principis ètics del model PIO

Sabem que l'ús de la tecnologia en general i de la IA en particular no és sempre positiva i desitjable o que pot afectar els drets humans així com l'estatus econòmic i molts altres aspectes de persones i grups que es veuen afectades, la qual cosa fa necessari establir un principi general de prudència (Steels i López de Mántaras, 2018). No obstant, per exercir-lo, és necessari que els principis ètics, que són vistos com a essencials en diferents revisions de la literatura (Jobin et al., 2019), siguin aplicats d'una manera pràctica i generalitzada. Si bé els primers principis ètics de la IA provenen del camp de la ciència ficció, tal i com ho demostren les Tres Lleis de la Robòtica contingudes en l'obra Runaround (Asimov, 1942), durant els últims anys, els principis ètics en la IA han passat de ser una qüestió filosòfica a convertir-se en una necessitat tangible. La proliferació dels *smartphones*, de les dades massives o *Big Data* i de les aplicacions d'IA, especialment de l'anomenat aprenentatge automàtic o *machine learning*, en tots els sectors industrials i també en els àmbits de la salut, de la provisió de serveis socials, de la justícia, del transport, de les finances o l'entreteniment, ha fet sorgir el debat per establir i aplicar aquells principis que donin sentit als usos ètics en les dades i els sistemes d'IA.

En aquest sentit, no és d'estranyar que l'augment de l'ús de tecnologies que inclouen sistemes d'IA hagi obert un ampli debat sobre l'impacte i les conseqüències del seu ús. De fet, algunes de les aplicacions actuals tant d'IA com els algorismes de presa de decisions automatitzades estant sent utilitzades en àrees considerades sensibles com són la vigilància o la policia predictiva. De manera més general, la IA també s'utilitza amb la finalitat d'obtenir informació en temps real sobre el comportament d'una multitud de persones, predir àrees de risc i modelar intervencions de polítiques públiques. Així doncs, a la vegada que la IA s'ha anat implantant cada vegada més en la nostra vida quotidiana, també ens hem adonat que els sistemes d'IA poden mantenir i, fins i tot, amplificar diferents biaixos negatius cap a diferents grups de persones, com les dones, les persones majors, les persones amb discapacitat i també cap a persones d'ètnies minoritàries, grups racialitzats o altres grups vulnerables. Com a conseqüència, una de les preguntes més urgents, especialment en el context de l'aprenentatge automàtic, és com evitar el biaixos que reprodueixen a través de les dades massives. L'existència de biaixos en l'ús de la IA no solament pot menystenir aquesta situació aparentment positiva de l'ús de la IA en la nostra vida quotidiana, sinó que pot acabar generant una falta de confiança en aquesta tecnologia, especialment entre les persones més afectades pels biaixos.

La IA és un ventall que aglutina moltes tecnologies relacionades, algunes de les quals estan tenint un progrés molt notable però accelerat. Per tant, hem de comprendre que de la mateixa manera que anteriors desenvolupaments tecnològics, als

beneficis de la utilització de sistemes d'IA, com poden ser l'automatització de tasques repetitives (algunes d'elles perilloses) o l'optimització i eficiència de processos, també sorgeixen una sèrie d'inconvenients que cal tenir en compte a l'hora de valorar els usos ètics de la IA. Per aquesta raó i amb l'objectiu d'explorar i avaluar el disseny, desenvolupament i implantació dels sistemes d'IA, l'OEIAC ha desenvolupat una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'IA que s'anomena model PIO (Principis, Indicadors i Observables)[3] i que està basada en 7 principis ètics fonamentals: (1) transparència i explicabilitat; (2) justícia i equitat; (3) seguretat i no maleficència; (4) responsabilitat i retiment de comptes; (5) privacitat; (6) autonomia; i (7) sostenibilitat. A continuació us detallem els principis i els seus aspectes més rellevants en relació a l'ús ètic de dades i sistemes d'IA.

2.1. Transparència i explicabilitat

Sens dubte, la transparència s'ha convertit en un dels principis més discutits dins dels debats sobre els usos ètics de la IA. D'acord amb l'informe presentat per Jobin et al (2019) sobre els valors globals de les estratègies d'ètica en IA, la transparència és el principi més repetit en la literatura actual (Floridi, 2019). La transparència es pot entendre de dues formes. D'una banda, fa referència al fet que la pròpia tecnologia de la IA sigui translúcida i que existeixi una obertura de la informació que s'utilitza. D'altra banda, existeix la transparència de les organitzacions que desenvolupen i/o utilitzen sistemes d'IA. Si bé la primera és discutida regular i directament, o en relació amb els processos requerits per a garantir-la en termes d'explicabilitat, comprensió i comunicació, la segona està en un segon pla i ha d'avançar per a garantir i protegir altres requisits com els drets humans fonamentals, la privacitat, la dignitat, l'autonomia i el benestar (UNI Global Union, 2017). Per això, les organitzacions que dissenyen i/o usen sistemes d'IA han de ser transparents sobre el seu objectiu inicial, i garantir que, amb una política de transparència, els usuaris puguin realitzar eleccions informades sobre l'intercanvi de les seves dades i l'ús de sistemes d'IA (ADMA, 2013).

Així doncs, la transparència és una condició fonamental per garantir la protecció dels drets humans fonamentals i els principis ètics i està estretament relacionada amb les mesures de retiment de comptes i, en definitiva, amb la fiabilitat dels mateixos sistemes d'IA. La preocupació per la no transparència ve donada pels riscos que un sistema d'IA pugui provocar i per la necessitat d'explicabilitat. Aquesta última es refereix a la comprensió i l'accés a la informació de les diferents parts que componen un algorisme i que es poden atribuir als resultats del sistema, ja que és essencial que els usuaris puguin qüestionar els resultats basant-se en informació senzilla i fàcil d'entendre (OECD, 2019).

Alguns estàndards com el IEEE (2022) estableixen diferents nivells de transparència

[3] Una part de l'acrònim PIO s'inspira en la definició d'entrada-sortida programada "Programmed Input/Output" dins de l'àmbit de transmissió de dades, mentre que l'altra està associada a l'expressió anglesa "Pass It On".

recomanats depenent del grau d'interacció amb el sistema. Principalment, això inclou el següent: que els usuaris que utilitzin un sistema de IA hagin de conèixer com funciona i per què produeix uns determinats resultats; que les persones que certifiquen els sistemes puguin conèixer el procés de presa de decisions; que les persones que investiguen des de diferents àmbits de coneixement siguin capaces de rastrejar els processos interns que condueixen a un mal funcionament o accident; i que les persones del públic general estiguin informades sobre com funciona un sistema de IA. Per tant, podríem dir que considerem que un sistema és transparent si les seves operacions no són opaques i són comprensibles per a les persones (European Parliament, 2017).

Encara que això pot suposar un problema, ja que la capacitat actual d'aprenentatge dels sistemes d'IA augmenta la dificultat d'explicació i pot significar que els propis dissenyadors no siguin capaços de comprendre el comportament i/o els resultats que proporciona un sistema d'IA, totes les organitzacions han de comprendre el funcionament i decisions aconseguides per aquestes tecnologies, sempre que sigui possible (Floridi et al., 2018). Evidentment, la capacitat de comprensió i explicació dels sistemes d'IA també es veu afectada pel cicle de vida de la pròpia eina d'IA que, moltes vegades, és externalitzada o venuda a tercers. Aquest procés, malgrat que inevitable, dificulta d'una banda, la retiment de comptes i, per un altre costat, la capacitat d'actuació i intervenció al llarg del propi cicle de vida. En aquest mateix sentit, les organitzacions haurien de ser capaces de comprendre com es va prendre una decisió i com la supervisió humana assegura que els possibles danys causats per l'anomenat black-boxing algorítmic són abordats i previnguts (IEEE, 2019). En aquest apartat es considera que en els àmbits d'alt risc (com a atenció de la salut, justícia i benestar) l'ús de les caixes negres de la IA hauria de reconsiderar-se per complet (AI Now Institute, 2017).

La transparència també suposa que totes les organitzacions han de documentar com els seus sistemes d'IA prenen unes certes decisions i han de poder reproduir-les per a una auditoria externa (SIIA, 2017). Si bé en alguns casos existeix una tensió entre rendiment i explicabilitat del sistema d'IA, aquesta tensió ha de documentar-se i estar clarament identificada (Cerna Collectif, 2018). Per aquest motiu, la transparència també depèn de la traçabilitat per a garantir que, si sorgeixen danys, es pugui rastrejar la causa (IEEE, 2017). Això significa que ha d'existir la disponibilitat de suficient informació detallada sobre les dades i accions d'un sistema d'IA perquè aquestes accions siguin reproduïbles i retrocedir posteriorment (OCDE, 2019). Per aquesta raó, és necessari dur a terme un seguiment de les dades utilitzades durant l'entrenament així com les condicions en les quals es van prendre i validar per, en definitiva, garantir la traçabilitat del sistema d'IA i oferir protocols de seguretat (Demiaux i Abdallah, 2017). Això és important ja que, encara que no utilitzem dades susceptibles de contenir biaixos, un sistema d'IA pot estar reconstruint valors esbiaixats i prenent decisions basats en ells. D'aquesta manera, és necessari parar esment a les dades d'entrenament i a les condicions mitjançant les quals s'han recollit aquestes

dades al llarg de tota l'operació del sistema. D'altra banda, en seleccionar aquestes dades s'ha de garantir la seva representativitat pensant en la diversitat dels grups d'usuaris.

Encara que interrelacionats, el principi de traçabilitat no és equivalent al de transparència. Per exemple, algunes guies com la proposta del European Law Institute (2022) posen l'accent en la necessitat d'informar a les persones afectades dels criteris i correlacions que han portat a la presa de decisions d'un sistema de IA, per no pas en la necessitat d'informar com s'han dut a terme aquestes decisions. En aquest sentit, és l'operador de la IA, la persona final responsable de la presa de decisions, la que ha d'estar en coneixement de les causes d'una decisió per a així poder auditar, controlar i/o monitorar el sistema d'IA, així com respondre en cas que sigui requerit per la persona afectada.

En definitiva, la transparència i explicabilitat fa referència a què el resultat d'un sistema d'IA es pugui explicar, interpretar i comunicar correctament. Es presenta com una forma de minimitzar riscos i també és fonamental per al compliment d'objectius socials com la participació, l'autoregulació i la confiança. En aquest sentit, els resultats i processos d'una intel·ligència artificial han de ser comprensibles i garantir la traçabilitat, especialment quan els resultats de la presa de decisions automatitzades afecten directament persones o col·lectius vulnerables, ja que els individus tenen dret a saber quan es pren una decisió sobre la base d'algorismes d'IA, i exigir o sol·licitar explicacions i informació a l'organisme pertinent, ja sigui aquest una empresa privada o institucions del sector públic. Per a això, segons la recomanació de la UNESCO (2021) cal trobar un equilibri just entre la transparència i la explicabilitat i altres principis com la seguretat i la protecció.

2.2. Justícia i equitat

La IA s'ha de desenvolupar perquè pugui utilitzar-se dins de la societat de manera justa i equitativa (NSTC, 2016). En aquest sentit, podem entendre que un sistema d'IA és just si no perjudica les persones i promou la igualtat dels individus respecte als seus drets, dignitat i llibertat per a prosperar (Tieto, 2018). Per a això és fonamental obtenir una major diversitat de les persones que treballen en el disseny i implantació de sistemes d'IA (Sage, 2017). En aquest sentit, les organitzacions no sols han de reduir l'exclusió, sinó que han de promoure la inclusió activa de dones i grups minoritaris en el desenvolupament i disseny de sistemes d'IA (Gilburt, 2019; WEF, 2018).

Encara que cal esperar que les persones que dissenyen els sistemes d'IA tenen els seus propis valors, han d'existir mecanismes de control perquè aquestes no deuen desenvolupin algorismes amb prejudicis històricament injustos (Latonero, 2018). També és necessari prendre més mesurades per a abordar danys sexistes i misògins, així com mitigar biaixos cognitius i estadístics per a garantir que les dades que utilitzen els sistemes d'IA no contenen errors i inexactituds, la qual cosa podria corrompre les respostes i decisions que

prenguin els sistemes d'IA (ICO, 2017).

Generalment, la discriminació i els resultats injustos derivats de l'ús de sistemes d'IA s'han convertit en un tema molt candent dins dels mitjans i cercles acadèmics (O'Neil, 2016), encara que la justícia algorítmica també fa referència a proporcionar recomanacions sobre com implementar els passos per a minimitzar aquests danys. Així doncs, per a prevenir accions nocives en el procés de presa de decisions, les organitzacions han d'assegurar-se que es recopilen dades de mostra precisos i representatius (UNDG, 2017). L'ús de la IA no ha de convertir-se en una eina més d'exclusió dins de la societat, i ha de prestar-se especial atenció a les persones infrarrepresentades i vulnerables, com les persones amb discapacitat, les minories ètniques, els nens i les persones del món no industrialitzat (AI HLEG, 2019).

Per a aconseguir dades d'alta qualitat és necessari que existeixin procediments regulars de verificació interns i externs amb els usuaris finals i grups d'interès que poden veure's afectats (PwC, 2019). Així doncs, s'espera que els actors implicats en el desenvolupament de sistemes d'IA implementin mecanismes adequats que permetin protegir i respectar els drets humans, prevenint l'aparició de biaixos algorítmics que puguin afectar les persones, especialment als grups més vulnerables. En el cas d'ocórrer una injustícia derivada de l'ús de sistemes d'IA, les diferents organitzacions haurien d'implementar mètodes per a revertir, remeiar i permetre una compensació justa a instàncies dels danys ocorreguts.

Tenint en compte que un dels objectius dels sistemes d'IA hauria de ser empoderar i beneficiar a les persones, és fonamental proporcionar igualtat d'oportunitats mentre es distribueixen les recompenses del seu ús d'una manera justa i equitativa (IEEE, 2019). Això suposa compartir els beneficis de les tecnologies d'IA a nivell local, nacional i internacional i tenir en compte les necessitats específiques de cada grup d'usuaris. Així doncs, és necessari abordar al llarg de tot el cicle de desenvolupament de la IA, la discriminació, l'accés al coneixement, així com la participació de les comunitats més vulnerables. Finalment, també cal subratllar que, atès que la utilització de sistemes d'IA pot suposar una substitució de persones per màquines, és important que existeixin polítiques justes i efectives de capacitació de la força laboral (COMEST/UNESCO, 2017). Les ramificacions de no fer-ho poden derivar en situacions injustes a nivell individual i col·lectiu, amb el que factors com l'accés a la tecnologia, la connectivitat i les competències dels individus, han de tenir-se en compte per a promoure la justícia i equitat.

Així doncs, la justícia i equitat fa referència a què un sistema d'IA pugui desplegar-se d'una manera justa i que les persones tinguin un tracte just o imparcial. Això inclou la prevenció, el seguiment o la mitigació de biaixos i discriminacions no desitjades, i també mecanismes d'apel·lació a decisions algorítmiques.

2.3. Seguretat i no maleficència

Els sistemes d'IA han de ser robustos i segurs durant tot el seu cicle de vida perquè no presentin riscos de seguretat ja que les conseqüències pels seus resultats o els accidents del seu mal ús poden afectar tant individus com a societats (OCDE, 2019). Per aquest motiu, les organitzacions han de garantir que la seguretat està integrada en l'arquitectura dels sistemes d'IA (The Public Voice, 2018). Primer, passant processos de control de qualitat i provar-se en el món real, de manera que es puguin reproduir escenaris reals abans, durant i després del desplegament (Algo.Rules, 2019; SAP, 2018). Segon, garantint una ciberseguretat efectiva perquè la seva IA estigui protegida contra atacs (Allistene, 2014). Les organitzacions també han de complir estrictes mesures de manejabilitat i control, i quan es troben vulnerabilitats o falles de disseny, aquestes s'han de divulgar per a ser resoltes (Internet Society, 2017).

Quan ens referim a la seguretat dels sistemes d'IA no sols fem referència a garantir que la tecnologia sigui segura des d'una perspectiva tècnica del seu funcionament (*security*), també ens referim al fet que els sistemes d'IA no infringeixin sobre els drets humans, avaluant els riscos de seguretat pública (*safety*) que sorgeixen de la seva implementació (The Public Voice, 2018). Així doncs, els sistemes d'IA han de dissenyar-se amb la intenció de no fer mal a les persones i testar-los regularment per a determinar que efectivament no resulta cap mal d'ells (ACM, 2017). Per a això, les organitzacions també han de rebre i incorporar l'assessorament de les autoritats legals i comitès d'ètica per a garantir que les dades de sistemes d'IA puguin recuperar-se, analitzar-se i utilitzar-se de manera que no perjudiquin les persones involucrades.

Els actors involucrats en el seu desenvolupament han de garantir la traçabilitat dels sistemes (IA de confiança) per a aplicar un enfocament sistemàtic en la gestió del risc durant cada fase. Per tant, els sistemes d'IA han de programar-se amb una condició de reversibilitat, que garanteix el control i seguretat del sistema. Dit d'una altra forma, hem de tenir la capacitat de desfer l'última acció o una seqüència d'accions per a poder retrocedir d'accions no desitjades i tornar a una etapa de seguretat (Clark, 2019). En aquest sentit, és important assenyalar que quan parlem de fiabilitat dins d'un sistema d'IA, ens referim al fet que quan estan sent utilitzats, es comporten d'acord amb les especificacions marcades per les persones dissenyadores. D'aquesta manera, en el cas que es produeixi un mal hauria de poder garantir-se l'aplicació de procediments d'avaluació de riscos i mesures perquè no torni a produir-se el mal.

En definitiva, la seguretat fa referència a la prevenció i protecció contra els danys i riscos que els sistemes d'IA poden causar. A aquest principi hem d'incloure el respecte, protecció i promoció de la dignitat humana, els drets humans i les llibertats fonamentals. La seguretat és un principi que ha de vetllar per la no maleficència, la innocuïtat i proporcionalitat dels sistemes d'IA. D'acord amb l'informe del High-Level Expert Group on AI, els sistemes d'IA

han de ser segurs, precisos, fiables i reproduïbles, perquè és l'única manera de garantir que es puguin minimitzar i prevenir els danys no intencionals (AI HLEG, 2019).

Per tant, la seguretat i no maleficiència respon a una crida general que els sistemes d'IA haurien de promoure el bé i mai haurien de causar danys previsibles o no intencionats. També fa referència a incloure mecanismes de seguretat per afrontar problemes com ara la ciberguerra o la pirateria maliciosa.

2.4. Responsabilitat i rendició de comptes

uan fem referència a la responsabilitat ens referim principalment a la responsabilitat moral, i aquesta fa referència tant a la gestió i ús responsable de les dades dels usuaris com a la responsabilitat dels actors involucrats en el desenvolupament i implantació d'un sistema d'IA. Així doncs, les organitzacions han de ser conscients dels problemes relacionats amb l'ús dades deficients i ser responsables si hi ha conseqüències perjudicials com a resultat d'això.

La responsabilitat s'ha erigit com un tema clau per a avançar en els usos ètics de la IA ja que hi ha el temor que les organitzacions puguin ofuscar la culpa o amagar-se de la seva responsabilitat en els sistemes autònoms o semi autònoms. La creixent autonomia relativa dels sistemes d'IA pot crear una seriosa d'incidències que poden derivar en l'anomenada "bretxa de responsabilitat", la qual cosa suposa que no és clar qui és el responsable. Per això, les organitzacions han de ser obertes i responsables per mitjà de l'auditoria, el seguiment i realització d'avaluacions d'impacte dels sistemes d'IA (ICDPPC, 2018).

Això significa que han d'establir mecanismes interns i externs que garanteixin la responsabilitat i el retiment de comptes. Per a això, és necessari que els sistemes d'IA siguin clars i íntegres en l'atribució de la responsabilitat legal, ja que ha de ser sempre possible atribuir la responsabilitat jurídica en qualsevol etapa del cicle de vida del sistema a persones físiques o entitats jurídiques. Aquest valor de retiment de comptes també fa referència a la supervisió humana del sistema. La decisió de cedir el control en contextos limitats haurà de continuar recaient en les persones, que seran les responsables dels possibles danys causats per la IA i aquesta culpa no pot atribuir-se als sistemes o eines d'IA (Algo.Rules, 2019).

En aquest mateix sentit, alguna proposta com la feta pel European Law Institute (2022) fa referència al principi de responsabilitat al fet que els operadors d'eines d'IA han de tenir en consideració l'impacte global de les decisions exercides per un sistema de presa de decisions automatitzat. És a dir, les possibles conseqüències que pugui tenir en els valors democràtics, els drets i llibertats de les persones, l'estabilitat del mercat, el medi ambient i la sostenibilitat o la cohesió social. En definitiva, es tracta d'una proposta a través de la

qual s'han de tenir en compte l'impacte potencial en els drets individuals i col·lectius i en l'entorn socioeconòmic.

De la mateixa manera, s'espera que els actors involucrats en el sistema d'IA siguin responsables del correcte funcionament del sistema i del respecte a tots els principis. En aquest cas, la traçabilitat és fonamental, en primer lloc, per a l'atribució de responsabilitat i pel retiment de comptes. És a dir, potencialment com a base dels procediments judicials, i en segon lloc per al diagnòstic i la correcció de disfuncions. Al mateix temps, la traçabilitat assegura que el funcionament d'un sistema es pugui explicar a partir del propi disseny del sistema d'IA, d'aquí la importància paral·lela de l'explicabilitat i de la creació d'escenaris potencials com una manera de bregar ex ante i ex post amb els danys potencials que puguin ocórrer (EGE, 2018; FATML, 2016).

En definitiva, la responsabilitat i retició de comptes fa referència a actuar amb integritat en les accions i decisions d'un sistema d'IA i també a la necessitat d'atribuir responsabilitat, incloent-hi en termes legals, a determinades persones com a responsables finals dels sistemes d'IA.

2.5. Privacitat

Salvaguardar la privacitat dels usuaris no és solament una condició indispensable per a la protecció de l'autonomia individual, sinó que permet l'autodesenvolupament, l'autorealització i la participació democràtica. Des que la Llei General de Protecció de Dades o RGPD va entrar en vigor en 2018, la privacitat ha estat un tema candent per a qualsevol que treballi en camps on s'estan utilitzant dades personals. La privacitat és vista com un valor a defensar relacionat amb la protecció de dades i el tractament d'aquests, així com la personalització de les dades, la transparència i la supervisió. Les aplicacions d'IA per si mateixes suposen un risc per a la privacitat a causa de les grans quantitats de dades que necessiten, de manera que, per a protegir la privacitat és necessari integrar-la dins del propi disseny de la IA. Així doncs, juntament amb el consentiment informat, els usuaris han de tenir control i accés a les dades emmagatzemades sobre ells (IEEE, 2019).

Cal subratllar que la privacitat també és percebuda com un dret que ha de ser protegit i que és essencial per a la protecció de la dignitat i l'autonomia de les persones. A més de garantir-se el ple dret de la privacitat i la protecció de dades, els sistemes d'IA han de garantir la governança de dades, assegurant l'accés legítim a aquests. Per aquesta raó, els desenvolupadors i organitzacions que utilitzen sistemes d'IA, han de col·locar la privacitat i les dades personals de l'usuari final en primer lloc, veient la privacitat com un dret humà (Latonero, 2018). A més, sempre que sigui possible, la recollida i l'ús de les dades personals ha de mantenir-se al mínim, tret que sigui completament necessari i rellevant (Datatilsynet, 2018). En aquest sentit, la Unió Europea ha començat a imposar mesures per a evitar l'ús

de sistemes considerats inacceptables tenint en compte la gran quantitat d'informació recollida així com l'abús i el risc que suposen per a la privacitat. Així, per exemple, l'ús de sistemes d'identificació biomètrica o reconeixement facial en l'espai públic utilitzats per a vigilància i que suposen un alt risc per a la població queden prohibits (European Comission, 2021).

Així doncs, la privacitat es presenta sovint en relació amb la protecció de dades, i també es relaciona amb la llibertat i confiança, sent un valor a mantenir i també com un dret a protegir.

2.6. Autonomia

Les organitzacions que utilitzen sistemes d'IA han d'assegurar-se que el "principi de l'Les organitzacions que utilitzen sistemes d'IA han d'assegurar-se que el "principi de l'autonomia de l'usuari ha de ser central a la funcionalitat del sistema" (High-Level Expert Group on AI, 2019: 16). Entenem per autonomia la llibertat d'expressió i apoderament dels usuaris enfront dels sistemes d'IA, eliminant la dependència dels seus usuaris enfront dels models de presa de decisions automatitzada. En les societats democràtiques l'autonomia es valora d'una forma particular ja que aquest principi està relacionat amb la protecció de la llibertat i la capacitat que la IA s'utilitzi sense destorbar-nos ni perjudicar-nos (EGE, 2018). En aquest sentit, qualsevol organització que utilitzi sistemes d'IA hauria d'assegurar-se que les persones donin el seu consentiment sobre com s'utilitzen les seves dades (Council of Europe, 2017) i, alhora, la IA no hauria de danyar les capacitats dels individus per a prendre decisions o manipular la seva autodeterminació (European Group on Ethics in Science and New Technologies, 2018). Això significa que els sistemes d'IA han de permetre l'apoderament de les persones, permetent-los prendre decisions informades i promovent els drets fonamentals.

L'autonomia d'un sistema depèn del grau de complexitat del seu entorn, que pot ser mesurat per la quantitat i varietat d'informació que conté així com de la complexitat de tasques que exerceixi i dels possibles estats del sistema. L'autonomia es troba promoguda per la transparència i la predictibilitat de la IA, la qual cosa vol dir no reduir les opcions i el coneixement als ciutadans, educar en IA a la població i complir amb els estàndards legals d'extracció de dades (Jobin, 2019).

Per a generar autonomia és necessari l'accés, la sensibilització i comprensió del públic respecte a les tecnologies d'IA, que hauria de promoure's a través de la capacitació de les competències digitals i la participació ciutadana. Aquestes mesures han de ser adoptades tenint en compte la diversitat lingüística, social i cultural existent, a fi de garantir una participació pública efectiva, de manera que tots els membres de la societat puguin adoptar decisions informades sobre la seva utilització dels sistemes d'IA i estiguin

protegits d'influències indegudes, particularment aquells que poden ser vulnerables a l'abús (Rathenau Institute, 2017).

Per tant, l'autonomia fa referència amb el control de l'usuari i amb la llibertat de triar a través de l'apoderament tecnològic, i també es refereix a la llibertat de retirar el consentiment.

2.7. Sostenibilitat

D'acord amb Vinuesa et al. (2020), la IA pot arribar a facilitar el 79% dels Objectius de Desenvolupament Sostenible, entenent la sostenibilitat com la capacitat de generar efectes positius i a llarg termini des de tres perspectives: la social, l'econòmica i la mediambiental.

D'una banda, el desenvolupament d'una IA sostenible socialment ve a assenyalar la necessitat de models de desenvolupament que permetin establir eines de comunicació i control entre els sistemes d'IA, els desenvolupadors i la societat. Això és perquè aquesta última és la que finalment viu amb major intensitat les conseqüències de l'aplicació de la IA, en termes de riscos i oportunitats. D'aquesta manera és essencial ser conscient de la complexitat existent d'interaccions i la creixent necessitat d'una legislació basada en l'ètica i els mecanismes d'avaluació i certificació de dades i sistemes d'IA.

Entre els aspectes que destaquen sobre el desenvolupament sostenible a nivell econòmic cal subratllar els efectes de l'automatització sobre les ocupacions i les desigualtats socioeconòmiques. Si bé la pèrdua d'ocupació es troba entre un dels majors temors sobre la implementació de la IA, la generació de noves ocupacions també es previsible però depèn d'una transició planificada on no tan sols intervé la demanda sinó també la preparació i la capacitació de persones amb determinades qualificacions relacionades amb la IA. Del contrari, és probable que hi hagi grups de població que vegin créixer la seva vulnerabilitat econòmica a través d'un accés al mercat laboral limitat. Per tant, l'impacte econòmic de la implementació, utilització o desenvolupament de IA ha de tenir en compte aquests aspectes sobre el desenvolupament sostenible a nivell econòmic.

Això significa que la promoció d'un desenvolupament sostenible de la IA és un valor transversal que podem trobar a través de l'alineació amb els altres principis ètics. Per aquesta raó, val la pena que la tractem com un principi independent però interconnectat. De fet, autors com Attard-Frost et al. (2022) subratllen que, per tal d'abordar la sostenibilitat de la IA, cal omplir àmpliament el seu abast disciplinari i incloure més dominis de coneixement als actuals i rellevants de caire tècnic i legal predominants.

D'altra banda, cal assenyalar que preocupa especialment el desenvolupament sostenible de la tecnologia i l'impacte mediambiental que produeix. En aquest sentit, les recomanacions de la UNESCO (2021) fan referència a la protecció i prosperitat del medi ambient i dels

sistemes, que ha de ser promogut al llarg del cicle de vida dels sistemes d'IA. Així doncs, les organitzacions públiques i privades han d'assegurar-se que respecten el medi ambient i incorporar els possibles impactes ambientals dins de la seva presa de decisions (Special Interest Group on Artificial Intelligence, 2018). En aquest sentit ha d'existir una adherència als recursos eficients, la promoció de l'energia renovable i la protecció de la biodiversitat en el disseny i ús de sistemes d'IA.

L'impacte ambiental i l'ús de recursos han de tenir importància en l'impacte del cicle de vida dels sistemes d'IA i, sens dubte, preocupa que les organitzacions usin la IA d'una manera que no sigui ambientalment conscient (SIIA, 2018). També se subratlla cada vegada més que, en situacions on hi ha mal ecològic causat per la IA, s'han de prendre mesures per a detenir-ho immediatament i identificar formes, si cap, d'usar la IA de manera no nociva. En cap cas ha d'usar-se la IA per a danyar la biodiversitat (UNI Global 2017) i, alhora, ha de protegir els entorns naturals (AI HLEG, 2019). Per això, l'ús de la IA ha de ser respectuós amb l'eficiència energètica i mitigar emissions de gasos d'efecte d'hivernacle, assegurant que la seva petjada ecològica sigui mínima i que tots els esforços siguin per a reduir els nivells d'emissió (Green Digital Working Group, 2016). Finalment, la IA ha de dissenyar-se de manera que garanteixi una energia i consum de recursos mínims i sense desaprofitament, especialment de materials escassos, i promogui l'eficiència dels mateixos (European Parliament, 2017).

En definitiva, la sostenibilitat respon a la necessitat d'un desplegament tecnològic de la IA respectuós amb la integritat ecològica i la justicia social al llarg del cicle de vida dels productes d'IA. Sovint s'utilitza per fer referència a tot el sistema sociotècnic d'IA, incloent-hi generació d'idees, formació, reajustament, implementació i governança.

2022

3 Definició, eines i model PIO (Principis, Indicadors i Observables)

En qualsevol procediment d'avaluació, la definició de l'objecte d'estudi és important ja que aquesta delimita els límits, també jurisdiccionals, de la mateixa avaluació. Aquesta delimitació conceptual és un tema clau per determinar quines tecnologies i tècniques cauen sota el paraigua de l'IA i, per tant, quina rellevància potencial tenen cadascun dels principis, indicadors i observables. En aquest sentit, la nostra proposta d'autoavaluació per a l'ús ètic de dades i sistemes d'IA, fem servir la definició actual de la IA recollida en l'actual Llibre Blanc de la Comissió Europea seguint les recomanacions de l'anomenat *High-Level Expert Group on Artificial Intelligence*:

"Els sistemes d'intel·ligència artificial (IA) són sistemes de programari (i possiblement també maquinari) dissenyats per humans que, donat un objectiu complex, actuen en la dimensió física o digital en percebre la seva entorn de través de l'adquisició de dades, interpretant les dades estructurades o no estructurats recopilats, raonant sobre el coneixement, o el processament de la informació, derivat d'aquestes dades i decidir la millor acció (és) a prendre per aconseguir l'objectiu dau. Els sistemes d'intel·ligència artificial poden usar regles simbòliques o aprendre un model numèric, i també poden adaptar el seu comportament analitzant com el medi ambient es veu afectat per les seves accions anteriors" (European Commission, 2020: 16).

Utilitzar aquesta definició propera de la Comissió Europea, ens permet moure'ns en una base mínima i operativa, que és un requisit previ per perfilar de forma pragmàtica l'abast dels sistemes d'IA, especialment en relació a qüestions ètiques i d'impacte social. Per tant, a través de la nostra proposta d'avaluació adoptem una definició que, encara que no sigui la més precisa possible i sigui susceptible de canviar, tingui en compte les raons per les quals ens preocupem per la IA avui dia. Evidentment, la importància d'una definició de la IA també radica en que cal una definició prou precisa per proporcionar la seguretat jurídica necessària.

Un cop adoptada la definició operativa, ens cal un seguit de directrius d'aplicació que puguin prevenir riscos no desitjables i, en definitiva, afavorir el disseny de sistemes d'IA que, adoptant normes ètiques, puguin ajudar-nos a avançar com a societat. En aquest sentit, la primera idea és implementar l'ètica dins dels sistemes d'IA com un dels principals objectius. Sabem que un dels punts de crítica cap a l'enfocament dominant de les directrius basades en principis és que poden simplificar en excés debats ètics, conduint-nos a una aparença de consens moral sense tenir en compte que són necessàries eines per al seu monitoratge, ja que les qüestions ètiques, especialment les més difícils, estan ocultes en els detalls de l'aplicació dels principis (Mittelstadt, 2019). És per aquesta raó que en la Declaració de Barcelona de 2017, Steels i López de Mántaras (2018) reclamen un principi de prudència ja que, d'una banda, l'aplicació de la IA basada en el coneixement requereix de la disponibilitat de suficient experiència humana i recursos per a una anàlisi

detallada i, d'altra banda, la IA basada en dades requereix de suficients dades d'alta qualitat i una elecció acurada d'algorismes i paràmetres per a cada cas. Així doncs, exercir la prudència necessària en l'aplicació de la IA per part dels responsables del seu desenvolupament apareix com a fonamental. Però les implicacions de la IA no són tan sol una qüestió de desenvolupament i validació interna, també tenen a veure amb una sèrie de requisits externs (Wagner, 2018) que tant organitzacions com institucions han de tenir en compte per a avançar cap als usos ètics de la IA:

- Participació externa, la qual cosa significa un compromís primerenc amb totes les parts interessades.
- Proporcionar un mecanisme de supervisió externa independent (no necessàriament públic).
- Assegurar procediments transparents de presa de decisions sobre el perquè es prenen unes determinades decisions.
- Desenvolupar una llista estable d'estàndards no arbitraris on la selecció d'uns certs valors i drets poden justificar-se de manera plausible.
- Assegurar que l'ètica no substitueixi als drets fonamentals ni als drets humans.
- Proporcionar una declaració clara sobre la relació entre els compromisos assumits i marcs legals o reguladors existents, en particular sobre el que succeeix quan els dos estan en conflicte.

Actualment existeixen diferents eines que vetllen pel compliment i l'alineació dels sistemes d'IA amb diferents principis ètics. En aquest sentit, d'acord amb Ada Lovelace Institute (Ada Lovelace Institute & DataKind, 2020) les eines més utilitzades són, d'una banda les auditories algorítmiques, i d'altra banda les avaluacions d'impacte algorítmic. Encara que totes dues fan referència a pràctiques que poden utilitzar-se per a avaluar sistemes algorítmics, cadascuna d'elles té una característica pròpia sobre el quan es realitza l'avaluació. Si bé les auditories algorítmiques s'enfoquen primordialment ex post, per determinar els possibles biaixos i riscos que es puguin produir una vegada el sistema d'IA ja s'ha implementat, les avaluacions d'impacte algorítmic tracten d'auditar els sistemes d'IA ex ante, per avaluar els possibles impactes d'un sistema d'IA abans que aquest sigui utilitzat. Evidentment, els dos tipus d'avaluació són compatibles i recomanables, especialment si tenim en compte que un sistema d'IA té un cicle de vida on poden aparèixer situacions i contextos no necessàriament anticipats.

En aquest sentit, tots aquests tipus d'eines d'avaluació, siguin ex ante, ex post o ambdues estan guanyant pes i protagonisme per determinar els usos ètics de les dades i sistemes d'IA ja que, històricament, les auditories que s'han utilitzat en el camp de l'enginyeria i la informàtica han estat focalitzades en uns pocs principis ètics com la seguretat i la no maleficència, i en determinar l'eficiència d'un algorisme i la seva adequació per resoldre un problema determinat. Així doncs, és una bona notícia que es comenci a generalitzar l'ús de les auditories o avaluacions per analitzar el disseny, ús i conseqüències de les dades i sistemes d'IA.

Però, evidentment, per tal que això funcioni és important tenir en compte que l'adequació dels valors ètics d'un sistema d'IA depenen en gran mesura del context cultural on s'apliqui la tecnologia. Això vol dir que la implementació i priorització d'uns certs valors dins de l'auditoria depèn del camp d'aplicació i el context cultural on operi el sistema d'IA. Així mateix, els múltiples factors de desenvolupament i implementació de la tecnologia poden influenciar el seu impacte. L'impacte social depèn no sols de la tecnologia, sinó dels objectius subjacents a un sistema i el lloc que ocupa la IA es troba dins de l'estructura d'una organització. En aquest sentit, hi ha dos punts importants a considerar. En primer lloc, que els valors es poden centrar en una ètica relativament individualista o de context singular. Per exemple, no són el mateix les consideracions en la salut pública que en la investigació mèdica. En segon lloc, cal tenir en compte que els usos i implementacions de la IA poden afectar individus que no són necessàriament clients o usuaris i, de fet, poden tenir un impacte en la societat en el seu conjunt. Per tant, en el model PIO també es tenen en compte preguntes a través de les quals s'apel·la a la responsabilitat directa i indirecta de les organitzacions.

Finalment, també és evident que la facilitat en l'ús dels models d'IA ètica pot tenir diferents significats depenent de les parts interessades o implicades. És a dir, els models d'implantació pràctica de la IA ètica requereixen d'eines que tinguin en compte els diferents rols de cada agent implicat en el sistema, ja siguin desenvolupadors, gestors, directors o usuaris. D'aquesta manera, totes les persones involucrades poden aportar informació rellevant sobre el seu funcionament i ús. Per això és essencial que en la consecució del model PIO a través de l'autoavaluació, hi prenguin part diferents persones de l'organització que es vol avaluar. Sabem que recollir informació a través d'aquestes eines no és senzill i que poden sorgir una sèrie de problemes, que van des de la naturalesa distribuïda dels reptes i oportunitats, la concentració i desequilibri de recursos o les possibles motivacions mixtes de diferents actors d'una mateixa organització. No obstant, el model PIO hauria de poder evidenciar les condicions i els aspectes organitzatius que formen un teló de fons sobre l'ús ètic de dades i sistemes d'IA. De fet, com qualsevol avaluació, es tracta de ser (més) conscients de les capacitats i límits d'un sistema tenint en compte les implicacions morals des de l'òptica organitzativa.

3.1. *Toolkits* i *checklists* ètics

Tot i que pot semblar que les consideracions ètiques al voltant de la IA són un tema nou ja fa temps que s'han proposat diversos recursos, especialment des del món acadèmic i d'algunes organitzacions sense ànim de lucre, que es poden dividir en els anomenats *toolkits* i *checklists*. Mentre els primers són eines que poden ajudar a fer que les aplicacions i els sistemes d'IA basats en dades per part d'una organització siguin més justos, més robusts o transparents, els segons són llistes de preguntes per determinar si un projecte o organització s'adequa o no als usos ètics d'un determinat context i requereix o no d'una revisió ètica en l'ús de dades i sistemes d'IA.

En tots dos casos, es tracta d'auditar els sistemes d'IA i les dades que s'utilitzen per anticipar i/o avaluar possibles impactes alhora que s'ajuda en la seva documentació i, és clar, en la presa de decisions sobre el seu ús i idoneïtat. Malgrat que la majoria d'aquestes eines són aspiracionals, també poden ser operacionals en accions concretes i, això, és especialment el cas en què principis ètics es pensen com a orientació per a persones que estan en la primera línia del desenvolupament de la IA. En aquest cas, podem dir que serveixen per observar d'una manera holística totes les etapes del cicle de vida del sistema d'IA, incloent-hi l'abast de possibles casos d'ús, els conjunts de dades, l'entrenament dels models d'aprenentatge automàtic, la seva implementació, etc. I, generalment, tendeixen a cobrir diverses categories de principis ètics com la privacitat, la robustesa del sistema, la seguretat i altres.

En aquest context, l'autoavaluació ètica és considerat un primer i important pas ja que suposa fer-nos un seguit de preguntes generals i específiques sobre el disseny, operacionalització i aplicació de sistemes d'IA, especialment respecte a l'ús de *machine learning*, que és un tipus d'intel·ligència artificial que permet que aplicacions o algorismes de programari millorin de manera automàtica a través de l'experiència i ús de dades històriques com a entrada per a predir nous valors de sortida d'acord (Cerna Collectif, 2018):

- Quin problema volem resoldre amb el sistema d'IA?
- El sistema d'IA només portarà a terme les tasques per a la qual ha estat dissenyat?
- Quines dades seleccionarem o utilitzarem perquè el sistema d'IA aprengui d'elles?
- Com podem avaluar a un sistema d'IA que aprèn a través d'un entrenament?
- Quines decisions poden ser o no delegades a un sistema d'IA?

- Quina informació s'ha de donar als usuaris sobre les capacitats d'un sistema d'IA?

- Qui és la persona responsable si el sistema d'IA no funciona correctament: el dissenyador, el propietari de les dades, el propietari del sistema o el propi sistema?

Fig. 1. Les cinc fases seqüencials d'un sistema d'IA (cicle de vida)

Generalment, aquestes qüestions que fan referència al cicle de vida d'un sistema d'IA es poden visualitzar seqüencialment en cinc etapes (Fig. 1) i, és clar, en cada una d'elles trobarem diferents aspectes que podem avaluar a través dels anomenats *toolkits* i *checklists*. Així doncs, en la primera etapa de predisseny és on s'enquadra el problema, la investigació i l'adquisició de dades. En aquest sentit, podem preguntar sobre els aspectes que condueixen a una determinada formulació d'un problema a resoldre així com identificar qui és el responsable de prendre les decisions i quant control hi ha sobre el procés de presa de decisions. Això permet un seguiment més evident de la responsabilitat en el desenvolupament del sistema d'IA abans de ser provats i examinats adequadament.

En la segona fase, que habitualment involucra científics de dades, enginyers i d'altres perfils similars, es focalitza en la part de l'enginyeria del sistema d'IA, la seva modelització i una primera avaluació. Així doncs, en aquesta part cal preguntar-se sobre la posada en pràctica de les assumpcions teòriques o operacionalització i, entre altres, els objectius d'optimització del sistema.

En la tercera fase, s'arriba a l'apartat de l'entrenament del sistema d'IA. En aquest sentit, cal tenir en consideració l'ajustament dels models (*over-/under-fitting*) en funció de les característiques seleccionades i avaluar els possibles biaixos, incloent-hi dimensions com el sexe i el gènere, l'ètnia i lloc de naixement, la classe socioeconòmica i el lloc de residència, entre altres. Evidentment, aquí s'ha de tenir en compte el context en què es desplegarà el sistema d'IA ja que utilitzar un model precís no mitiga automàticament cap problema de biaix. Un exemple és utilitzar un sistema d'IA que indica ser un 95% precís en una població jove però que pot ser inútil si s'utilitza amb una població d'edat avançada. És per aquesta raó que en aquesta fase és imprescindible comptar amb la participació de persones que puguin participar activament en qüestionar el procés ja que això pot traduir millor el context social i prevenir problemes associats a les anomenades "fuites de destinació".

La quarta fase consisteix en avaluar el desplegament havent tingut en compte el context, la seva possible adaptació i les limitacions pràctiques. En aquesta fase és on acostuma a sorgir qualsevol biaix no identificat prèviament tenint en compte que els usuaris comencen a interactuar amb la tecnologia.

Finalment, en la darrera fase cal realitzar un seguiment després del desplegament que consisteix en evitar o minimitzar a través de la interpretació dels resultats alguns problemes centrals com la sobregeneralització, les fal·làcies de correlació, els biaixos automatitzats i els biaixos de confirmació. Si bé la majoria de biaixos després del desplegament no són intencionats, la interacció general del públic amb el sistema d'IA fa que es pugui exposar qualsevol problema. Malgrat que aquest no hauria de ser el cas i que aquests problemes s'han de tractar en les 2 etapes anteriors, la tendència a tractar la fase de desplegament i monitorització com a fases de proves fa que augmentin aquests problemes. El principal contratemps d'aquesta situació és que ja no estem davant d'una mitigació dels biaixos sinó d'una reacció retardada que pot haver tingut conseqüències molt negatives.

L'ús dels *toolkits* i *checklists* és una forma efectiva per avaluar diferents fases i els processos que les componen ja que, malgrat poden simplificar principis i aspectes complexos (Mandaio et al., 2020), aquests permeten abordar i validar qüestions que no sempre són tractades en la formulació, construcció, entrenament, desplegament i seguiment de qualsevol sistema d'IA. Per tant, el seu ús és molt aconsellable ja que d'una forma generalitzada podria aconseguir un impacte positiu, especialment quan s'identifiquen en les diferents fases del cicle de vida d'un sistema d'IA, les parts responsables i la possible exposició de com entren el biaixos en el processos. No menys important és el fet que suposa compartir i implicar diferents membres del context social en què es desplegarà el sistema d'IA, i aquesta implicació pot conduir a una comprensió més elaborada i més profunda de les implicacions socials de la IA. De fet, aquest format d'auditoria està molt influenciat pels *checklists* realitzats en el desenvolupament de programari, on s'utilitzen per a estandarditzar processos com la revisió de codi. No obstant, això no suposa que tot hagi de focalitzar-se en solucions purament tècniques ja que, en lloc de deixar-ho a unes poques persones seleccionades en el procés de disseny, cal tenir en compte que la IA, com qualsevol tecnologia, és més efectiva i acceptada quan és representativa de moltes persones i no pas quan es vol fer a través dels ulls d'uns pocs.

En l'actualitat hi ha nombrosos *toolkits* que han estat desenvolupats per organitzacions privades i públiques, i que estan enfocats a diferents consideracions ètiques com la responsabilitat, i a problemes recurrents com els biaixos. En aquest sentit, us presentem alguns dels més coneguts:

Audit-AI[4], és una eina per mesurar i mitigar els efectes dels patrons discriminatoris en

[4] Pymetrics (2020, 29 juliol). AI Audit. GitHub, https://github.com/pymetrics/audit-ai (consultada el 23 de juny de 2022).

les dades d'entrenament i les prediccions fetes per algorismes d'aprenentatge automàtic entrenats per a processos de decisió socialment sensibles.

Playing with AI Fairness[5], és una eina que els investigadors i dissenyadors de la iniciativa PAIR (People and AI Research) de Google van crear com a recurs pels desenvolupadors de sistemes d'aprenentatge automàtic. Específicament, l'ús de l'eina *What-If* pretén ajudar a desenvolupadors i usuaris en una de les preguntes més difícils i complexes plantejades pels sistemes d'IA: què volem dir quan ens referim a un sistema just?

Ethics & Algorithms Toolkit[6], és una eina que permet als diferents governs i al seu personal entendre i gestionar l'ús dels sistemes d'IA. Aquesta respon principalment a la pressió creixent del públic, els mitjans de comunicació i les institucions acadèmiques perquè siguin més transparents i responsables sobre el seu ús.

AI Explainability 360[7], es tracta d'un conjunt d'eines de codi obert de IBM per ajudar a comprendre com els models d'aprenentatge automàtic prediuen etiquetes per diversos mitjans tenint en compte el seu cicle de vida.

PwC's Responsible AI[8], es un conjunt d'eines i processos personalitzables dissenyats per ajudar a les organitzacions a utilitzar la IA d'una manera responsable, des de l'estratègia fins a l'execució.

InterpretML[9], és un paquet de codi obert de Microsoft que incorpora tècniques d'interpretació d'aprenentatge automàtic d'última generació. Amb aquest paquet es pretén millorar la interpretació i explicació dels sistemes d'IA considerats com a caixes negres.

Responsible Innovation: A Best Practices Toolkit[10], és un conjunt d'eines de Microsoft que proporcionen als desenvolupadors un conjunt de pràctiques per anticipar i abordar els possibles impactes negatius de la tecnologia sobre les persones.

Bias and Fairness Audit Toolkit[11], és un conjunt d'eines d'auditoria de biaixos de codi obert dissenyades per Aequitas - Center for Data Science and Public Policy – University of Chicago, per a desenvolupadors, analistes i responsables polítics per auditar els models d'aprenentatge automàtic per discriminació i biaix, i ajudar a prendre decisions informades

[5] Weinberger, D. (2018) Playing with AI Fairness. What-If Tool, https://pair-code.github.io/what-if-tool/ai-fairness.html (consultada el 23 de juny de 2022).

[6] Anderson, D., Bonaguro, J., McKinney, M., Nicklin, A., & Wiseman, J. (2018). Ethics & Algorithms Toolkit (beta). Ethics Toolkit, de https://ethicstoolkit.ai/ (consultada el 23 de juny de 2022).

[7] IBM Research (2019) AI Explainability 360. IBM Research Trusted AI, http://aix360.mybluemix.net/(consultada el 23 de juny de 2022).

[8] PWC (2019) Responsible AI Toolkit, https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html (consultada el 23 de juny de 2022).

[9] Microsoft (2021, 23 setembre). InterpretM. GitHub, https://github.com/interpretml/interpret (consultada el 23 de juny de 2022).

[10] Microsoft (2022, 6 maig). Responsible Innovation toolkit: Azure Application Architecture Guide. Microsoft Docs, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/ (consultada el 23 de juny de 2022).

[11] Center for Data Science and Public Policy, University of Chicago. (2018). Bias and Fairness Audit Toolkit. Aequitas: Bias & Fairness Audit, http://aequitas.dssg.io/(consultada el 23 de juny de 2022).

i més equitatives sobre el desenvolupament i el desplegament d'eines predictives d'avaluació de riscos.

Fairlearn[12], és un paquet Python de Microsoft que permet als desenvolupadors de sistemes d'IA avaluar l'equitat del seu sistema i mitigar qualsevol problema d'injustícia observat. Fairlearn conté algorismes de mitigació, així com un giny (Jupyter widget) per a l'avaluació del model. A més del codi font, aquest dipòsit també conté quaderns de Jupyter amb exemples d'ús de Fairlearn.

AI Verify[13], és un marc i un conjunt d'eines d'autoavaluació de codi obert per a proves de governança d'IA desenvolupat per les autoritats de Singapur per avançar en una IA responsable a través de la promoció de la transparència. En aquest sentit, desenvolupadors i propietaris de sistemes d'IA poden verificar el rendiment dels seus sistemes amb un conjunt de principis mitjançant proves estandarditzades.

Tenint en compte la varietat de sistemes d'IA, aplicacions i enfocaments metodològics i, és clar, fonts de dades, existeixen iniciatives que contenen un *checklist* o llista de verificació i que tenen l'avantatge d'abordar qüestions menys tècniques o específiques però, alhora, capaces de reconèixer els temes i aspectes claus de les dades i sistemes d'IA per tal que siguin tractades com un flux de treball estàndard. Generalment, es tracta de llistes, formularis o qüestionaris que permeten estructurar i considerar les principals idees sobre les implicacions ètiques dels projectes en què una organització està treballant. En aquest sentit, els *checklists* permeten avançar i animar als diferents agents d'un ecosistema, incloent-hi persones que es dediquen a al desenvolupament, disseny, emprenedoria, administració, gestió o educació, a comprometre's amb els usos ètics de les dades i els sistemes d'IA. És amb aquest objectiu que trobem, principalment, la presència clara d'iniciatives governamentals i també d'algunes organitzacions privades.

A continuació us presentem algunes de les més rellevants:

Algorithmic Impact Assessment Tool del Govern del Canadà[14], permet determinar el nivell d'impacte d'un sistema de decisió automatitzat a partir de 48 preguntes de risc i 33 de mitigació. En aquest checklist, les puntuacions d'avaluació es basen en molts factors, com ara el disseny dels sistemes, l'algorisme, el tipus de decisió, l'impacte i les dades.

The Artificial Intelligence (AI) Ethics Framework del Govern d'Australia[15], permet identificar processos i solucions sobre com s'han aplicat 8 principis ètics i estan dissenyats

[12] Fairnlearn (2021, 7 de juliol) FairLearn. GitHub, https://github.com/fairlearn/fairlearn (consultada el 23 de juny de 2022).

[13] Personal Data Protection Comission Singapore (PDPC) (2022, 25 de maig) Singapore's Approach to AI Governance, https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework (consultada el 23 de juny de 2022).

[14] Government of Canada. (2022, 25 de febrer). Algorithmic Impact Assessment, de https://open.canada.ca/aia-eia-js/?lang=en (consultada el 23 de juny de 2022).

[15] Australian Government: Department of industry, Science, Energy and Resources. (2019). Australia's Artificial Intelligence Ethics Framework, https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework (consultada el 23 de juny de 2022).

per demanar a les organitzacions que considerin l'impacte de l'ús de sistemes habilitats per IA. També proporciona comentaris sobre àrees que necessiten una major orientació a partir d'aprenentatges i consells.

AI System Ethics Self-Assessment Tool del Govern de Dubai[16], també ha creat la seva pròpia eina d'autoavaluació que està pensada per ajudar a identificar els possibles problemes ètics que poden sorgir al llarg del procés de desenvolupament, incloses les etapes inicials de la idea fins al manteniment del sistema quan estigui en ple funcionament. També ofereix algunes idees sobre quin tipus de mesures de mitigació es podrien introduir.

Algorithmic Transparency Standard del Govern del Regne Unit[17], és un estàndard de transparència algorítmica per als departaments governamentals i els organismes del sector públic que ofereix una guia pràctica per analitzar la conveniència d'utilitzar un servei d'IA enfront d'un repte concret. Aquesta avaluació, permet als desenvolupadors avaluar la necessitat d'utilitzar un sistema d'IA, on s'està utilitzant, qui és el responsable, etc.

Data Ethics Framework del Govern del Regne Unit[18]**,** és una guia d'orientació per a organitzacions del sector públic sobre com utilitzar les dades de manera adequada i responsable a l'hora de planificar, implementar i avaluar una nova política o servei. Si bé és un marc ètic de les dades i no pas dels sistemes d'IA, és igualment útil donada la seva relació per guiar el disseny de l'ús adequat de les dades al sector públic.

Understanding Artificial Intelligence Ethics and Safety del Alan Turing Institute[19], és una guia per al disseny responsable i la implementació de sistemes d'intel·ligència artificial en el sector públic. Aquesta guia repassa tot el cicle de vida del projecte d'IA i està integrat en un marc integral d'IA responsable.

Algorithmic Accountability Policy del AI Now Institute[20]**,** es podria definir com una guia de polítiques de responsabilitat algorítmica que té com a objectiu proporcionar als defensors legals i polítics una comprensió bàsica de l'ús governamental dels algorismes, incloent-hi un desglossament dels conceptes i preguntes clau que poden sorgir quan tracten amb sistemes d'IA, especialment en termes de transparència i rendició de comptes.

Autoevalución ética de IA para actores del ecosistema emprendedor del Banco Interamericano de Desarrollo[21]**,** és una iniciativa que promou l'ús ètic i responsable de la IA a través d'una autoavaluació. L'objectiu és ajudar els emprenedors a millorar el

[16] Digital Dubai Office (2020) AI System Ethics Self-Assesment Tool, https://www.digitaldubai.ae/self-assessment (consultada el 23 de juny de 2022).

[17] Central Digital and Data Office. (2022, 1 de juny). Algorithmic Transparency Standard. United Kingdom Government, https://www.gov.uk/government/collections/algorithmic-transparency-standard (consultada el 23 de juny de 2022).

[18] Central Digital and Data Office. (2020, 16 de setembre). Data Ethics Framework. United Kingdom Government, https://www.gov.uk/government/publications/data-ethics-framework (consultada el 23 de juny de 2022).

[19] Leslie, D. (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

[20] AI Now Institute (2018, octubre). Algorithmic Accountability Policy Toolkit (N.o 01), https://ainowinstitute.org/aap-toolkit.pdf

[21] Rosales Torres, C. S., Buenadicha Sánchez, C., & Narita, T. (2021, maig) Autoevaluación ética de IA para actores del ecosistema emprendedor: Guía de aplicación. Banco Interamericano de Desarrollo, http://dx.doi.org/10.18235/0003269

desenvolupament de la IA a la vegada que ajuda a identificar les àrees d'atenció principals per prevenir errors, biaixos, discriminacions i exclusions resultants del desplegament tecnològic.

Empowering AI Leadership del World Economic Forum[22], és un recurs per a consells d'administració que consta de 13 mòduls destinats a fomentar el coneixement i actuar davant de les possibles preocupacions ètiques en l'ús dels sistemes d'IA. Els primers vuit mòduls se centren en la supervisió de l'estratègia d'IA i la responsabilitat tenint en compte la marca, la competència, els clients, el model operatiu, les persones i la cultura, la tecnologia, la ciberseguretat i el desenvolupament sostenible. Cinc mòduls més cobreixen temes addicionals de supervisió com l'ètica, el govern, el risc, l'auditoria i responsabilitats del consell d'administració.

Tal com esmentem el principi del document, el *High-Level Expert Group on Artificial Intelligence* també ha desenvolupat la seva pròpia avaluació titulada *Assessment List for Trustworthy Artificial Intelligence*[23] (ALTAI) i que consisteix en una eina pràctica que ajuda organitzacions, empreses i desenvolupadors a autoavaluar la confiabilitat dels seus sistemes d'IA. Aquest concepte d'IA de confiança es troba definit en *Ethics Guidelines for Trustworthy Artificial Intelligence* (AI HLEG, 2019) i està basat en els següents requeriments:

- Agència humana i supervisió.
- Robustesa tècnica i seguretat.
- Privacitat i governança de dades.
- Transparència.
- Diversitat, no discriminació i equitat.
- Benestar ambiental i social.
- Responsabilitat.

Cadascun d'aquests requeriments han estat traduïts en una detallada llista d'avaluació. Aquest sistema ha estat testat en una àmplia consulta pública amb la comunitat europea d'IA i la seva avaluació, feta en un format complet i comprensible, està adaptat al context europeu, la qual cosa és un factor positiu i negatiu alhora ja que a vegades pot resultar massa específic amb aquest context. Generalment, els *checklists* ètics permeten contrastar l'existència o no a través de respostes binàries si existeix o no una acció o preocupació

[22] World Economic Forum (2020) Empowering AI Leadership, https://www.weforum.org/projects/ai-board-leadership-toolkit (consultada el 23 de juny de 2022).

[23] AI HLEG (2020) The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, European Commission, Brussels, https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence (consultada l'11 de març de 2022).

ètica respecte al desenvolupament i implantació d'un sistema d'IA.

Més enllà d'aquests *checklists* de caire governamental (alguns supranacionals), també trobem alguns de menys ambiciosos però igualment rellevants com els desenvolupats per *Deon*[24], que permeten afegir una llista de verificació d'aspectes ètics a projectes relacionats amb la ciència de dades o *Ethical OS*[25], que és una llista de verificació entre les quals trobem característiques com l'avaluació de 8 zones de risc per ajudar a identificar les àrees emergents de risc i danys socials més crítiques, 14 escenaris per conversar i imaginar impactes a llarg termini, i 7 estratègies de futur per ajudar a prendre mesures ètiques. En la mateixa línia, existeix l'eina del *Markkula Center for Applied Ethics* (Santa Clara University)[26], que ajuda a implementar la reflexió, la deliberació i el judici ètic en els fluxos de treball d'enginyeria i disseny de la indústria tecnològica.

3.2. Certificacions ètiques

Un altre dels models d'avaluació són les certificacions ètiques, la idea de les quals consisteix en una identificació, insígnia o segell de qualitat que acredita la conformitat d'un sistema d'IA de manera externa. Es poden distingir diferents tipus de certificació que són útils per a la certificació de sistemes d'IA, incloent-hi una certificació de producte o una forma mixta de producte i procés de certificació. Generalment, la certificació del producte és una verificació del compliment de la garantia de propietats del producte, i aquesta hauria d'iniciar-se en una fase inicial o d'especificació del sistema. La certificació de procés pot ser una alternativa o complement a la certificació de producte ja que s'examina la qualitat de fabricació, desenvolupament, així com el procés general i particular de la solució d'IA. Aquesta última serveix per reflexionar sobre els processos a provar i, en determinades circumstàncies, també es poden dur a terme a càrrec del fabricant o del propi operador. La certificació de processos és important per prevenir possibles mal funcionaments i proporcionar una salvaguarda en termes de responsabilitat.

Dit això, una certificació depèn, en primer lloc, sobre el risc que pot constituir un sistema d'IA per a les persones i també sobre els béns legals. En segon lloc, les certificacions tenen en compte una gamma de possibles intervencions per part de les persones en un context determinat d'aplicació, sovint indicant nivells de risc (mínims i màxims) associats a una possible necessitat de regulació. En aquest context, el Llibre Blanc de la IA de la Comissió Europea[27], constitueix un primer punt de partida per a una certificació reeixida de sistemes d'IA, abans no hi hagi normes ni estàndards reconeguts. La proposta de marc regulador de la IA realitzada per la Comissió Europea[28], on s'estableixen unes normes harmonitzades i

[24] Deon (s. f.) An ethics checklist for data scientists, https://deon.drivendata.org/ (consultada el 23 de juny de 2022).

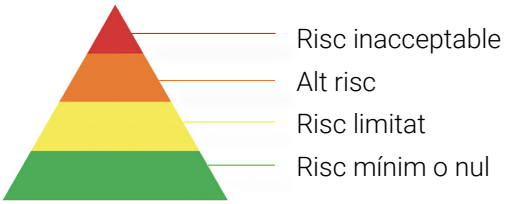
[25] Institute For The Future (2020).Ethical OS Toolkit, https://ethicalos.org/(consultada el 23 de juny de 2022).

[26] Vallor, S., Green, B. & Raicu , I. (2018) Ethics in Technology Practice. The Markkula Center for Applied Ethics at Santa Clara University, https://www.scu.edu/ethics/.(consultada el 23 de juny de 2022).

[27] European Commission (2020). White Paper on Artificial Intelligence: A European approach to excellence and trust (COM(2020) 65). Brussels: European Commission, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (consultada el 27 d'agost de 2021).

[28] Madiega, T. (2021)Artificial intelligence act, EPRS: European Parliamentary Research Service, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf (consultada el 23 de juny de 2022).

s'aborden les preocupacions ètiques i de drets humans també constitueix una certificació ètica a través de 4 nivells de risc:



En el primer grup de risc inacceptable, trobem tots els sistemes d'IA que es consideren una amenaça clara per a la seguretat, els mitjans de vida i els drets de les persones i volen ser prohibits d'acord amb aquesta certificació. Aquest inclouen des de la puntuació social dels governs fins a les joguines que utilitzen assistència de veu que fomenten comportaments perillosos. Més concretament, la Comissió Europea proposa prohibir completament els sistemes d'IA que:

- Manipulen persones mitjançant tècniques subliminals o explotar la fragilitat d'individus vulnerables, i potencialment podria danyar la persona manipulada o una tercera persona.
- Serveixen per a finalitats generals de puntuació social, si es realitza per autoritats públiques.
- S'utilitzen per executar sistemes d'identificació biomètrica remota en temps real en espais d'accés públic amb finalitats d'aplicació de la llei.

Dins el segon grup, on s'identifiquen els sistemes d'IA d'alt risc, s'han inclòs:

- Infraestructures crítiques (per exemple, transport), que poden posar en risc la vida i la salut dels ciutadans.
- Formació educativa o professional, que pot determinar l'accés a l'educació i al curs professional de la vida d'algú (per exemple, la puntuació dels exàmens).
- Components de seguretat dels productes (per exemple, aplicació d'IA en cirurgia assistida per robot).
- Ocupació, gestió de treballadors i accés a l'autoocupació (per exemple, programari de classificació de currículums per als procediments de contractació).

- Serveis públics i privats essencials (per exemple, la puntuació de crèdit que denega als ciutadans l'oportunitat d'obtenir un préstec).

- Aplicació de la llei que pugui interferir amb els drets fonamentals de les persones (per exemple, avaluació de la fiabilitat de les proves).

- Gestió de la migració, l'asil i el control de fronteres (per exemple, verificació de l'autenticitat dels documents de viatge).

- Administració de justícia i processos democràtics (per exemple, aplicació de la llei a un conjunt concret de fets).

Els sistemes d'IA considerats d'alt risc estaran subjectes a obligacions estrictes abans que es puguin posar al mercat, i aquestes inclouen:

- Sistemes adequats d'avaluació i mitigació de riscos.

- Alta qualitat dels conjunts de dades que alimenten el sistema per minimitzar els riscos i els resultats discriminatoris.

- Registre de l'activitat per garantir la traçabilitat dels resultats.

- Documentació detallada que proporcioni tota la informació necessària sobre el sistema i la seva finalitat perquè les autoritats avaluïn el seu compliment.

- Informació clara i adequada per a l'usuari.

- Mesures adequades de supervisió humana per minimitzar el risc.

- Alt nivell de robustesa, seguretat i precisió.

Tots els sistemes d'identificació biomètrica remota es consideren d'alt risc i estan subjectes a requisits estrictes. En principi, està prohibit l'ús de la identificació biomètrica remota en espais d'accés públic amb finalitats d'aplicació de la llei. Es defineixen i regulen de manera estricta les excepcions limitades, com ara quan cal buscar un nen desaparegut, prevenir una amenaça terrorista específica i imminent o detectar, localitzar, identificar o perseguir un autor o sospitós d'un delicte greu. Aquest ús està subjecte a l'autorització d'un òrgan judicial o d'un altre òrgan independent i als límits adequats de temps, abast geogràfic i bases de dades consultades.

En el tercer grup de risc limitat, trobem els sistemes d'IA amb obligacions específiques de transparència. Quan utilitzen sistemes d'IA com ara *chatbots*, els usuaris han de ser conscients que estan interactuant amb una màquina perquè puguin prendre una decisió informada per continuar o fer un pas enrere. Finalment, hi ha un quart grup on la proposta permet l'ús gratuït d'IA per considerar-se de risc mínim. Això inclou aplicacions com ara videojocs amb IA o filtres de correu brossa. La gran majoria dels sistemes d'IA que s'utilitzen actualment a la UE entren en aquesta categoria.

Tenint en compte els nivells de risc proposats per la Comissió Europea, s'han desenvolupat models de certificació com el VCIO del *AI Ethics Impact Group* (AIEI Group)[29]. Aquest model, les sigles del qual volen dir Valors, Criteris, Indicadors, Observables, persegueix fer practicables, comparables i mesurables una sèrie de principis ètics. El model VCIO proposa un sistema universal d'avaluació per a aplicacions d'IA basat en un sistema de puntuació ètic. En aquest sentit, la norma general recomanada pel grup per a tractar la IA ètica serà la d'aplicar uns criteris d'avaluació i uns indicadors que puguin ser observables. Per valorar el compliment dels principis ètics, AIEI Group proposa l'establiment de sis valors clau que serveixin de barem i que són la transparència, la responsabilitat, la privacitat, la justícia, la confiabilitat i la sostenibilitat ambiental. Quant a la decisió de quin nivell d'etiqueta hauria de ser considerat com èticament acceptable, hem d'assenyalar que aquest podria variar depenent del sector on s'apliqui. No és el mateix un sistema d'IA utilitzat en un procés industrial on potser està subjecte a un menor nivell de transparència que el mateix sistema aplicat a un procediment mèdic que inclogui el processament de dades personals.

Aquest model proposa una sèrie de criteris que poden ser utilitzats per a identificar la intensitat d'un mal potencial que pugui resultar de l'ús d'un algorisme. A més, el model té en compte el grau de dependència que poden arribar a desenvolupar els individus en utilitzar un sistema de presa de decisions automatitzat. Els revisors han de considerar, per exemple, si la decisió automatitzada ha de ser revisada i verificada per un humà, si els individus poden triar en quina mesura estan subjectes a la presa de decisions automatitzada, o si existeixen procediments per a corregir decisions falses o nocives que podria prendre un sistema d'intel·ligència artificial. Els autors arriben a suggerir que, en el cas que el mal potencial sigui baix, un sistema de valoració ètica de la IA podria arribar a ser innecessari. Però a mesura que la dependència o el riscos augmentin, les qualificacions dels valors avaluats mitjançant el model VCIO, com la transparència, la confiabilitat o la justícia, podrien resultar de gran importància i requerir una revisió addicional. En els casos més extrems, per exemple, en les armes automàtiques quan el mal potencial és molt alt, els autors proposen que els components de *machine learning* no siguin admesos, o almenys, haurien d'ajustar-se per a aconseguir un nivell de classificació ètica permissible.

La certificació ètica de VCIO permet als desenvolupadors respondre a diferents preguntes

[29] AIEI Group (2020) AI Ethics Impact Group: From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, https://www.ai-ethics-impact.org/en (consultada el 27 de agosto de 2021).

catalogades com a "observables". En funció de les respostes es genera una nota que indica el grau de seguretat o compliment del valor. Per exemple, un determinat sistema d'IA pot tenir una qualificació d'A en transparència, C en responsabilitat, etc. Donat que la sensibilitat ètica d'un algorisme ve definida pel seu context d'utilització, el model VCIO també estima una matriu de risc en cinc classes que permet determinar fins a quin punt el sistema requereix de regulació. L'eix vertical determina el grau de dependència d'una persona afectada per la decisió d'una IA i l'eix horitzontal marca el grau de risc depenent dels seus danys potencials. Així es determina el resultat de la matriu entre les diferents classes:

Classe 0: la intensitat del mal i l'exposició de la decisió és tan baixa que no requereix una qualificació del sistema ètic.

Classe 1: si la intensitat del mal potencial i l'exposició a la decisió excedeixen un cert llindar, un organisme regulador ha de fer demandes inicials per a la qualificació ètica.

Classe 2: a causa de la intensitat del mal potencial i l'exposició de l'usuari, les dades d'entrada han de divulgar-se per complet i la informació sobre la qualitat del sistema s'ha de revisar. S'ha de minimitzar els danys si les decisions del sistema d'IA tenen un impacte negatiu significatiu sobre els individus o grups.

Classe 3: existeix cert risc de mal potencial de les decisions de la IA, s'han de reduir els riscos tant com sigui possible, és a dir, identificar i evitar qualsevol forma en la qual es puguin prendre decisions adverses.

Classe 4: el sistema d'IA té un potencial de mal total o tan alt que no han d'usar-se components d'aprenentatge automàtic en absolut (per exemple, sistemes d'armes autònomes).

És previsible que un cop identificada la necessitat d'una certificació en base al nivells de risc, les certificacions es vincularan a normes, estàndards, procediments de prova i regulacions específiques de la indústria com, per exemple, normes ISO, DIN o de la coneguda IEEE, que ha desenvolupat recentment la seva certificació ètica anomenada *IEEE CertifAIEd*. Aquesta certificació verifica el compliment del sistema amb una sèrie de criteris ètics (transparència, retiment de comptes, aparició de biaixos i privacitat) amb l'objectiu de fomentar la confiança envers els sistemes d'IA. Específicament, la certificació *IEEE CertifAIEd* (IEEE, 2021)[30] vol ajudar a les organitzacions a:

- Afirmar el seu compromís de defensar els valors humans, la dignitat, el benestar i de respectar, protegir i preservar els drets humans fonamentals.

[30] IEEE (2021) IEEE CertifAIEd, https://engagestandards.ieee.org/ieeecertifaied.html (consultada el 23 de juny de 2022)

- Transmetre la capacitat d'aquesta organització per a complir amb la transparència, la responsabilitat, la reducció del biaix algorítmic i els requisits de privacitat aplicables estipulats en els criteris apropiats per a fomentar la confiança i facilitar l'adopció i l'ús de productes i serveis d'IA.

- Augmentar la confiança en les empreses públiques i privades que desitgen obtenir els beneficis de la certificació d'ètica de la IA en absència o com a complement de les regulacions àmpliament acceptades i aplicades per a la IA, al mateix temps que mitiga els riscos, les responsabilitats i els impactes adversos en la seva reputació i participació de mercat.

Sovint, els criteris de prova que es volen aplicar es poden dividir en termes de caràcter vinculant en el marc d'una certificació de criteris mínims, el que vol dir que s'han de complir i comprovar segons la sol·licitud respectiva i el context. Juntament amb aquests criteris d'una certificació de mínims, existeixen els criteris addicionals que van més enllà i que permeten una certificació augmentada o plus. Aquests criteris són de gran importància per a un desenvolupament positiu i orientat als valors d'IA fiable i per anar més enllà dels requisits mínims, que generalment serveixen per evitar perills evidents i immediats. Per exemple, la plataforma alemanya Lernende Systeme[31], proposa els següents criteris mínims:

- Transparència, traçabilitat, verificabilitat i rendició de comptes.

- Tenir seguretat (funcional) i fiabilitat del producte.

- Evitar efectes conseqüents no desitjats (en altres sistemes, persones i el medi ambient).

- Vetllar per l'equitat en el sentit d'igualtat i no discriminació.

- Aconseguir protecció de la intimitat i la personalitat.

- Autodeterminació, inclosa la transparència sobre l'ús del sistema d'IA i el paper de l'ésser humà en el procés de presa de decisions.

[31] Lernende Systeme. (s. f.). Lernende Systeme: Publications, https://www.plattform-lernende-systeme.de/publications.html (consultada el 23 de juny de 2022)

3.3. La proposta i metodologia del model d'autoavaluació PIO

El model PIO ha de permetre a organitzacions públiques i privades poder adoptar una reflexivitat crítica però pragmàtica sobre el disseny, desenvolupament i implementació de dades i sistemes d'IA tenint en compte dues qüestions bàsiques:

- Esteu dissenyant, desenvolupant o implementant un sistema d'IA que s'utilitzarà per prendre decisions o donar suport a aquestes d'una manera que pot tenir un impacte significatiu (positiu o negatiu) en les persones (incloent-hi els grups vulnerables), el medi ambient o la societat?

- Esteu segurs de com pot afectar el vostre sistema d'IA tant als vostres usuaris i/o clients com a la vostra organització?

En el cas de respondre sí en la primera pregunta i no en la segona, el model PIO us pot ajudar a planificar millor els resultats, des d'una perspectiva organitzativa, sobre l'ús ètic de dades i sistemes d'IA. En aquest sentit, es tracta d'un model que es sustenta en la idea d'ús responsable de la tecnologia a través d'un millor coneixement i aplicació de principis ètics per tal d'identificar accions adequades o inadequades i oportunitats de millora. Si bé aquest model PIO està dirigit a persones d'organitzacions públiques i privades que actualment utilitzen sistemes d'IA o tenen previst fer-ho en un futur no molt llunyà, aquest model també està enfocat a qualsevol persona que tingui un interès general o particular en l'ús i implementació de sistemes d'IA i en la seva millora com a usuàries i/o clients.

A continuació detallem proposta i metodologia del *Model PIO (Principis, Indicadors, Observables): Una proposta d'autoavaluació organitzativa sobre l'ús ètic de dades i sistemes d'intel·ligència artificial*. Com hem vist fins ara, aquest és un treball que no s'entén sense la revisió documental prèvia, i en la que s'ha inclòs literatura acadèmica, documentació tècnica així com literatura grisa i recursos web de diferents organitzacions públiques i privades. Aquesta proposta, també es basa en l'anàlisi de documents i projectes de llei relacionats amb l'ús de la IA, principalment en el context europeu. Així doncs, juntament amb els continguts dels anomenats *toolkits*, *checklists* i les certificacions ètiques anteriorment revisats, els indicadors observables del model PIO han estat definits sobre la base de la següent documentació de referència: *Ethics Guidelines for Trustworthy AI* (European Commission, 2019), *Draft text of the Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2021), *Declaració de Barcelona* (IIIA CSIC, 2017), *AI Ethics Impact Group: From Principles to Practice* (AIEI Group, 2020), *Technical methods for regulatory inspection of algorithmic systems in social media platforms* (Ada Lovelace Institute, 2021), *AI systems classification framework at the OECD* (OECD, 2022), *AI and Data Protection Risk Mitigation And Management Toolkit* (ICO, 2021).

Malgrat que les consideracions ètiques al voltant de la IA poden ser molt específiques, complexes i veritables trencaclosques segons l'aplicació de la tecnologia i el context, la proposta i metodologia del model PIO d'autoavaluació parteix d'un sentit general a com funcionen aquestes tecnologies des d'un enfocament d'ètica aplicada de resolució de problemes. Així doncs, el model PIO proposa un desglossament de conceptes i preguntes clau relacionades amb els 7 principis ètics inclosos en el model que poden sorgir en tractar qualsevol tema per part de persones que estan dissenyant, desenvolupant o implementant sistemes d'IA. Aïllant els 7 principis ètics en blocs, el model PIO s'organitza en una sèrie de preguntes a les persones que conformen una organització sobre l'ús ètic de dades i sistemes d'IA, per comprovar el compliment d'aquests principis. Per fer-ho, el model PIO combina aspectes dels *toolkits*, *checklists* i certificacions ètiques, si bé el formulari o llista de verificació amb els indicadors observables és el component principal. Tal i com s'ha assenyalat al principi del treball, la raó principal d'utilitzar aquest tipus de proposta és que es construeix al voltant de la senzillesa i d'una pregunta clau i efectiva que qualsevol persona que desenvolupa, gestiona o dirigeix un projecte on hi ha dades i sistemes d'IA pot entendre fàcilment: Ho hem fet? A més, a diferència de la majoria de toolkits que es concentren en un o dos principis o aspectes tècnics d'avaluació, la proposta del model PIO cobreix una gran majoria dels aspectes que es tracten en la literatura especialitzada i en cadascun dels principis ètics revisats. D'aquesta manera, les organitzacions poden estar més segures sobre què necessiten saber i aplicar segons la documentació de referència en el context europeu i, per tant, no es sorprenguin sobre la manca d'aspectes d'avaluació, ans al contrari (si bé això mai es pot descartar). En aquest sentit, es pot subratllar que la proposta del model PIO no deixa gaire espai per moure's en l'ambivalència, i permet una anàlisi ràpida i/o completa sobre els usos ètics de les dades i els sistemes d'IA en termes organitzatius. Per descomptat, és possible que qualsevol persona que fa l'autoavaluació prengui dreceres i ho faci d'una manera que pot ser inadequada. No obstant, el model PIO està dissenyat per tal que, com a mínim, quedi palès quina és la posició i actuació envers cadascun dels principis ètics. És clar, si un projecte no inclou la interacció física amb les persones o no inclou dades sobre persones o dades que representen poblacions de persones, és probable que no calgui cap revisió o avaluació a través del model PIO, tret que hi hagi una preocupació particular que es vulgui tenir en compte.

El model PIO no ha estat com una llista de verificació obligatòria per a les organitzacions. De fet, no es tracta d'una legislació o reglament, sinó d'un model que les organitzacions, de forma voluntària, poden utilitzar a l'hora de dissenyar, desenvolupar, integrar o implementar usos ètics en les dades i els sistemes d'IA. Malgrat aquesta característica de voluntarietat, la intenció és que el model PIO sigui aspiracional i complementi, no pas substitueixi, les normatives i pràctiques d'IA responsable existents. En aquest sentit, el seu impacte pot ser molt beneficiós ja que té la capacitat de:

- Poder influir de manera positiva en els resultats de la IA i en tots els processos associats des d'un principi en el seu disseny fins al seu desplegament.

- Generar confiança pública envers els productes i/o serveis d'una organització que implementa el model PIO.

- Fomentar la fidelitat d'usuaris i consumidors en organitzacions que utilitzen el model PIO.

- Garantir que les persones es beneficiïn de l'ús ètic de les dades i dels sistemes d'IA.

El model PIO ofereix dues opcions per a les organitzacions: una autoavaluació ràpida de posició i una autoavaluació de posició completa amb recomanacions. Totes dues avaluacions (ràpida i completa) han de servir per tal de conèixer com i fins a quin punt les organitzacions han pres consideració dels 7 principis ètics del nostre sistema d'avaluació PIO sobre: (1) transparència i explicabilitat; (2) justícia i equitat; (3) seguretat i no maleficència; (4) responsabilitat i retiment de comptes; (5) privacitat; (6) autonomia; i (7) sostenibilitat. Considerem aquestes avaluacions com un pas actualment ineludible i fonamental per a millorar la presa de consciència generalitzada de principis ètics per a l'ús de sistemes d'IA i la capacitat de proporcionar informació formativa específica per a avançar en la seva aplicació.

El sistema d'avaluació estarà organitzat en quatre fases (Fig.2). En primer lloc, les organitzacions hauran de respondre a les preguntes seleccionades estratègicament per a cada principi (Fase 1), de manera que amb les respostes s'obtindrà una puntuació d'adequació amb diferents aspectes dels principis ètics depenent de les respostes obtingudes (Fase 2). En aquest sentit, abans de portar a terme l'autoavaluació serà important escollir la persona o persones responsables dins de les organitzacions públiques o privades que utilitzen sistemes d'IA per realitzar, des d'un punt de vista competencial, una autoavaluació amb la màxima informació disponible. Per fer-ho, es recolliran respostes binàries a 118 preguntes sobre cada un dels 7 principis anteriorment descrits, de manera que es recolliran dos valors possibles en les respostes (Sí/No | 1/0).

L'avantatge d'utilitzar respostes binàries és que minimitza la no resposta i facilita l'obtenció d'informació de forma ràpida i eficient. Aquest tipus de resposta binària s'utilitza sovint en mesures de qualitat, especialment en ciències conductuals, biològiques i socials on només hi ha dos valors possibles, com ara sí o no, vertader o fals. A més, és una aproximació que moltes persones programadores estan familiaritzades en diversos procediments estadístics. Amb la informació obtinguda de l'autoavaluació s'obtindran mètriques per cada un dels 7 principis avaluats i amb elles perfils observables de cada

una de les organitzacions que portin a terme l'autoavaluació (Fase 3). Un cop obtingudes les mètriques corresponents, es procedirà a l'elaboració d'una insígnia de qualitat que serà la imatge gràfica del model PIO representant l'adequació de les dades i del sistema d'IA que s'ha autoavaluat segons els 7 principis ètics. Per realitzar aquest procés s'utilitzarà un gradient de color d'acord que representarà els resultats obtinguts en el qüestionari d'autoavaluació (Fase 4).

Tal i com s'ha comentat anteriorment, es podrà realitzar una autoavaluació ràpida de posició i una autoavaluació de posició completa amb recomanacions. Aquesta última requereix la justificació de les respostes afirmatives, ja que això permetrà a l'equip d'OEIAC, juntament amb les persones col·laboradores externes, analitzar les respostes i poder proporcionar feedback a l'organització que l'hagi realitzat si així ho desitja. Cal tenir en compte que, a partir dels principis, indicadors i observables, es pot proporcionar una descripció de l'orientació i adequació que pren un sistema d'IA no sols per a les persones que el desenvolupen, sinó també per als seus usuaris. D'aquesta manera, a més d'obtenir unes mètriques específiques i una insígnia de qualitat, les organitzacions que hagin realitzat l'autoavaluació completa rebran recomanacions específiques de millora.

La insígnia de qualitat és una imatge gràfica que es basa en les mètriques obtingudes sobre l'orientació i adequació del sistema d'IA als diferents principis ètics d'acord amb el model PIO. En aquest sentit, en el seu interior se situen els diferents principis (Fig. 3) amb un gradient de color d'acord amb les mètriques obtingudes (Fig. 4).

Així doncs, a manera de resum aquestes són les fases que inclou cada model d'autoavaluació:

Autoavaluació ràpida: FASE 1+ FASE 2 + FASE 3 + FASE 4
Autoavaluació completa: FASE 1+ FASE 2 + FASE 3 + FASE 4 + Feedback OEIAC

Fig. 2. Fases d'execució del model PIO

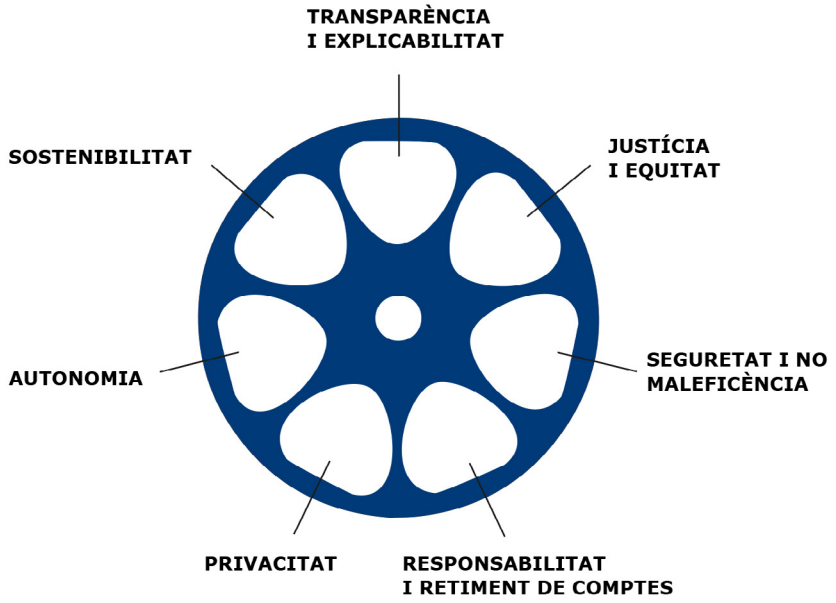
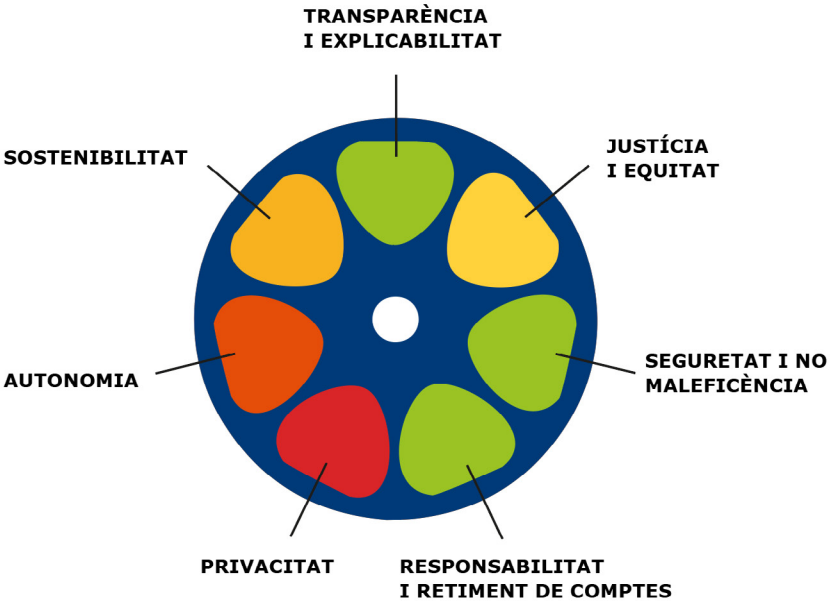


Fig. 3. Distribució dels principis del model PIO



Transparència i explicabilitat: 100%

Justícia i equitat: 80%

Seguretat i no maleficència: 100%

Responsabilitat i retiment de comptes: 100%

Privacitat: 0%

Autonomia: 20%

Sostenibilitat: 60%

Fig. 4. Exemple visual de resultats després de l'avaluació PIO

3.4. Continguts del qüestionari d'autoavaluació del model PIO

Sens dubte, les persones que conformen les organitzacions desenvolupadores i/o usuàries de sistemes d'IA tenen ara discussions que mai no haurien tingut lloc fa una dècada. No obstant, moltes vegades aquestes discussions no acaben de materialitzar-se processos d'avaluació i, encara més, en l'adopció i aplicació de decisions que permetin un ús més ètic de les dades i dels sistemes d'IA. A continuació, s'exposen les preguntes-indicadors que formen part del sistema d'avaluació del model PIO tenint en compte els 7 principis ètics: (1) transparència i explicabilitat; (2) justícia i equitat; (3) seguretat i no maleficència; (4) responsabilitat i retiment de comptes; (5) privacitat; (6) autonomia; i (7) sostenibilitat.

Les principals característiques dels 7 principis:

Transparència i explicabilitat: els sistemes d'IA han de ser transparents i explicables perquè la gent pugui entendre quan la IA els afecta de manera significativa i esbrinar quan un sistema d'IA s'hi interacciona.

Justícia i equitat: els sistemes d'IA han de ser inclusius i accessibles, i no han d'implicar ni provocar una discriminació injusta contra persones, comunitats o grups.

Seguretat i no maleficència: els sistemes d'IA han de funcionar de manera fiable d'acord amb el propòsit previst.

Responsabilitat i retiment de comptes: les persones responsables de les diferents fases del cicle de vida del sistema d'IA han de ser identificables i responsables dels resultats dels sistemes d'IA, i s'hauria d'habilitar la supervisió humana dels sistemes d'IA.

Privacitat: els sistemes d'IA han de respectar i defensar els drets de privadesa i la protecció de dades, i garantir la seguretat de les dades.

Autonomia: els sistemes d'IA han de respectar els drets humans, la diversitat i l'autonomia de les persones.

Sostenibilitat: els sistemes d'IA haurien de beneficiar les persones, la societat i el medi ambient.

Cadascun dels principis conté diferents aspectes sobre els quals les organitzacions han de proporcionar informació a través de les seves respostes i justificació si es realitza l'autoavaluació de posició completa amb recomanacions. En la següent taula 1 hi ha un

sumari dels principis i nombre de preguntes per cada principi. En total hi ha 70 preguntes i cadascuna d'elles representa un indicador observable.

Taula 1. Sumari de principis i nombre de preguntes

Indicador	Nombre total de preguntes observables
Transparència i explicabilitat	**16**
- Anàlisi preliminar de l'ús del sistema	5
- Explicabilitat	5
- Traçabilitat i representatibitat	3
- Protocols de comunicació	3
Justícia i equitat	**10**
- Inclusió i diversitat	6
- Identificació i mitigació de biaixos	4
Seguretat i no maleficència	**11**
- Funcionament del sistema	3
- Traçabilitat per a la gestió del risc	4
- Mecanismes de seguretat	2
- Protecció enfront d'atacs	2
Responsabilitat i rendició de comptes	**9**
- Control	3

- Retiment de comptes	3
- Figures de protecció	3
Privacitat	**10**
- Protecció de la privacitat	6
- Governança de dades	4
Autonomia	**6**
- Impacte en les capacitats dels usuaris	4
- Percepció del sistema	2
Sostenibilitat	**8**
- Desenvolupament sostenible	4
- Impacte mediambiental	4
TOTAL	**70**

Transparència i explicabilitat

Hi hauria d'haver transparència i divulgació responsable perquè les persones puguin entendre quan la IA els afecta de manera significativa i esbrinar quan un sistema d'IA s'està relacionant amb ells. En aquest sentit, la pregunta guia és la següent: **Si el vostre sistema d'IA funciona tal i com es pretén, es pot comprendre i explicar el seu funcionament?**

1. Anàlisi preliminar de l'ús del sistema			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	El procés de generació de resultats del vostre sistema d'IA està ben documentat i és reproduïble per si cal descobrir problemes en el futur?	Justificar resposta (500 caràcters max.)	ICO/ALTAI /OEIAC
1.2	Coneixeu amb precisió per a què s'utilitza el vostre sistema d'IA?		ICO/OEIAC
1.3	La resta de persones implicades, incloent-hi els usuaris, coneixen per a què s'utilitza el vostre sistema d'IA?		UNESCO/ OEIAC
1.4	Heu tingut en compte les capacitats o competències de tots els possibles usuaris, incloent-hi grups vulnerables, amb problemes d'accessibilitat i/o necessitats especials?		ALTAI/ UNESCO/ OEIAC
1.5	Sabeu si el vostre sistema d'IA utilitza el model més simple i intel·ligible en favor de la transparència?		ALTAI/AIEIG/ OEIAC

2. Explicabilitat			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Sou capaços d'entendre i explicar com el vostre sistema d'IA arriba a un resultat determinat?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
2.2	Sou coneixedors de les possibles limitacions del vostre sistema d'IA i les heu comunicat a tots els actors implicats, incloent-hi els usuaris?		ALTAI/OEIAC
2.3	Heu donat informació a les persones que utilitzen el vostre sistema d'IA sobre l'intercanvi de dades, els resultats previstos i com exercir els seus drets?		ICO/OEIAC
2.4	Considereu que heu establert un nivell de transparència i explicabilitat que no posa en risc la privacitat de les persones?		UNESCO/ OEIAC
2.5	Considereu que heu establert un nivell de transparència i explicabilitat que no posa en risc la seguretat del vostre sistema d'IA?		UNESCO/ OEIAC

3. Traçabilitat i representativitat			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
3.1	Heu documentat els tipus de dades, la seva procedència i els canvis efectuats en cada fase del cicle de vida del vostre sistema d'IA?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
3.2	Per tal de protegir les dades de persones menors, heu tingut en compte salvaguardes com la traçabilitat de les dades per verificar l'edat de les persones?		ALTAI/OEIAC

3.3	En el cas d'utilitzar dades demogràfiques, heu validat la seva representativitat tenint en compte el context més proper a través de la informació mostral o poblacional d'una organització d'estadística?		ICO/OEIAC

4. Protocols de comunicació

Nº	Indicador	Observable: Sí (1) /No (0)	Ref.
4.1	Heu establert eines de comunicació que permeten als usuaris del vostre sistema d'IA realitzar comentaris sobre el seu funcionament?	Justificar resposta (500 caràcters max.)	ALTAI/AIEIG/ UNESCO/ICO/ OEIAC
4.2.	Heu consultat amb d'altres organitzacions que hagin dissenyat i/o implementat un sistema d'IA semblant per conèixer antecedents i experiències?		ALTAI/OEIAC
4.3	Heu establert una política de participació de diferents grups d'interès (stakeholders), incloent-hi usuaris finals i/o receptors del vostre sistema d'IA?		ALTAI/AIEIG/ OEIAC

Justícia i equitat

Al llarg del seu cicle de vida, els sistemes d'IA han de ser inclusius i accessibles, i no haurien d'implicar ni provocar una discriminació injusta contra persones, comunitats o grups. En aquest sentit, la pregunta guia és la següent: **Quin és l'impacte individual i social del funcionament del seu sistema d'IA?**

1. Inclusió i diversitat			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	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?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
1.2	Dins del vostre equip de treball, s'inclouen dones que garanteixin la diversitat d'opinions dins del procés de disseny i implementació del sistema d'IA?		ALTAI/ICO/ OEIAC
1.3	Dins del vostre equip de treball s'inclouen persones de diversos orígens o cultures que garanteixin la diversitat d'opinions dins del procés de disseny i implementació del sistema d'IA?		ALTAI/ICO/ OEIAC
1.4	Dins del vostre equip de treball s'inclouen persones de diferents disciplines de coneixement més enllà de la informàtica i l'enginyeria?		ALTAI/ICO/ OEIAC
1.5	Us heu assegurat que el vostre sistema d'IA no utilitza variables o "proxies" que poden ser injustament discriminatòries?		UNESCO/ OEIAC

1.6	Heu establert mesures d'inclusió i equitat com provar els resultats del vostre sistema d'IA per a diferents grups afectats, per exemple provant per taxes d'error dispars?		UNESCO/ OEIAC

2. Identificació i mitigació de biaixos			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Heu realitzat una avaluació d'impacte dels possibles biaixos del vostre sistema d'IA per saber si podria reforçar discriminacions, estereotips o prejudicis envers persones vulnerables (e.g. persones amb discapacitat, minories ètniques o individus d'edat avançada)?	Justificar resposta (500 caràcters max.)	AIEIG/OEIAC
2.2	Heu implementat mesures que permetin mitigar els possibles biaixos estadístics en el vostre sistema d'IA, siguin originats en les dades d'entrenament o a posteriori?		ALTAI/OEIAC
2.3	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?		ALTAI/OEIAC
2.4	Heu adoptat mesures que permetin mitigar possibles biaixos cognitius en la fase de disseny o implementació del vostre sistema d'IA (e.g. el biaix de confirmació)?		ALTAI/OEIAC

Seguretat i no maleficència

Al llarg del seu cicle de vida, els sistemes d'IA haurien de funcionar de manera fiable d'acord amb el propòsit previst. En aquest sentit, la pregunta guia és la següent: **El vostre sistema d'IA funciona tal i com es pretén?**

1. Funcionament del sistema			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	Heu establert mecanismes de supervisió i control (intern o extern) per assegurar-vos que el vostre sistema d'IA funciona tal i com es pretén?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
1.2	Coneixeu els possibles comportaments i escenaris que el vostre sistema d'IA mai hauria de transgredir per garantir la seguretat i preservar la no maleficència?		ALTAI/OEIAC
1.3	Les persones que han dissenyat el vostre sistema d'IA han proporcionat orientació sobre com interpretar i respondre a les mètriques generades pel propi sistema?		ALTAI/OEIAC

2. Traçabilitat per a la gestió del risc			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Heu establert o adoptat un protocol general de traçabilitat de les dades per a la gestió del risc del vostre sistema d'IA?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
2.2	Heu avaluat que, en cap cas, les dades que recolliu, genereu o utilitzeu pel funcionament del vostre sistema d'IA no poden utilitzar-se per discriminar persones de forma negativa?		ALTAI/OEIAC
2.3	Heu avaluat els possibles riscs o danys que provocaria que el vostre sistema d'IA fes prediccions inexactes en els escenaris previstos?		ALTAI/OEIAC
2.4	Heu considerat el possibles riscs o danys que provocaria que el vostre sistema d'IA s'utilitzés per a altres usos més enllà dels previstos?		ALTAI/OEIAC

3. Mecanismes de seguretat			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
3.1	Heu establert protocols de seguretat (interns o externs) per detectar possibles problemes d'execució i/o autonomia del vostre sistema d'IA?	Justificar resposta (500 caràcters max.)	ALTAI/DBCN/ OEIAC
3.2	Heu establert mesures per examinar la precisió del vostre sistema d'IA de forma continua per conèixer si aquest està fent una quantitat inacceptable de prediccions inexactes?		ALTAI/OEIAC

4. Protecció enfront d'atacs			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
4.1	Heu avaluat els riscos i la vulnerabilitat del vostre sistema d'IA per garantir la seva seguretat, integritat i resiliència davant de diferents formes d'atac?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
4.2	Heu portat a terme una avaluació de protecció que garanteixi que els conjunts de dades que utilitza el vostre sistema d'IA no han estat prèviament compromesos o piratejats?		ALTAI/OEIAC

Responsabilitat i retiment de comptes

Els responsables de les diferents fases del cicle de vida del sistema d'IA haurien de ser identificables i responsables dels resultats dels sistemes d'IA, i s'hauria d'habilitar la supervisió humana dels sistemes d'IA. En aquest sentit, la pregunta guia és la següent: **Qui és responsable si el vostre sistema d'IA, especialment si deixa de funcionar correctament?**

1. Control			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	Us heu assegurat que les persones involucrades en el control tenen una comprensió adequada de la presa de decisions que efectua el vostre sistema d'IA?	Justificar resposta (500 caràcters max.)	ICO/OEIAC
1.2	Heu definit clarament els límits del nivell de control o participació humana dins del vostre sistema d'IA?		ALTAI/OEIAC
1.3	Poden els usuaris controlar fins a quin punt estan exposats i revocar el consentiment en qualsevol moment sense que això comprometi el rendiment del sistema d'IA?		AIEIG/OEIAC

2. Retiment de comptes			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Heu determinat qui és legalment responsable de les accions i/o decisions que pren el vostre sistema d'IA durant tot el seu cicle de vida?	Justificar resposta (500 caràcters max.)	Mörch et al (2019)/UNESCO/OEAC

2.2	Heu avaluat de quina manera el vostre sistema d'IA influeix en els processos de presa de decisions de la vostra organització?		ALTAI/OEIAC
2.3	Heu establert mecanismes de revisió periòdica sobre responsabilitat i retiment de comptes amb els diferents agents involucrats en el desenvolupament i ús del vostre sistema d'IA?		AIEIG/ALTAI/ OEIAC

3. Figures de protecció			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
3.1	Heu determinat una o més figures de protecció envers l'usuari del vostre sistema d'IA com, per exemple, la figura de defensor/a del poble?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
3.2	Heu establert formes de reparació o compensació en cas que ocorri algun mal o impacte advers de responsabilitat a través de pòlisses d'assegurances, mitjans monetaris adequats i/o altres formes de compensació?		ALTAI/OEIAC
3.3	Heu pres mesures per facilitar la demostració del compliment de la legislació vigent incloent-hi, per exemple, la Llei de Protecció de Dades (GDPR)?		ICO/OEIAC

Privacitat

Al llarg del seu cicle de vida, els sistemes d'IA haurien de respectar i defensar els drets de privadesa i la protecció de dades, i garantir la seguretat de les dades. En aquest sentit, la pregunta guia és la següent: **Es garanteix la privacitat de les persones amb l'entrenament i funcionament del vostre sistema d'IA?**

1. Protecció de la privacitat			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	Heu establert mecanismes de protecció de dades per tal que les dades de caràcter personal siguin tractades amb el grau de protecció legalment requerit i així evitar-ne el mal ús, alteració, pèrdua o sostracció per part de tercers?	Justificar resposta (500 caràcters max.)	ICO/OEIAC
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?		ICO/OEIAC
1.3	Heu establert un protocol per obtenir el consentiment abans d'iniciar qualsevol processament o anàlisi a través del vostre sistema d'IA?		ICO/OEIAC
1.4	La informació sobre el consentiment i la seva revocació és comprensible i accessible per tots els usuaris, incloent aquells que pertanyen a grups vulnerables, persones amb problemes d'accessibilitat o necessitats especials?		AIEIG/OEIAC

1.5	Heu establert processos per mantenir o millorar la privacitat i la protecció de dades al llarg del cicle de vida del vostre sistema d'IA (com el xifratge, l'anonimització o l'agregació)?		AIEIG/ALTAI/ OEIAC
1.6	Heu establert un mecanisme a través del qual una persona pot sol·licitar que la seva informació personal sigui eliminada?		AIEIG/ALTAI/ OEIAC

2. Governança de dades			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Heu establert un registre de la informació d'accés a les dades (personals o no) per identificar clarament quan, qui i amb quin propòsit s'ha accedit a les dades?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
2.2	Us heu assegurat que les persones que realitzen el processament i gestió de dades compten amb les competències necessàries i es mantenen al dia amb els nous desenvolupament per a comprendre i dur a terme la política de protecció de dades?		ALTAI/OEIAC
2.3	Heu establert els procediments per assegurar que els conjunts de dades siguin els adients i rellevants pel correcte rendiment i resultat previst del model?		Pereira/ OEIAC
2.4	Heu documentat el procés de recopilació i preparació de les dades pel vostre sistema d'IA?		Pereira/ OEIAC

Autonomia

Al llarg del seu cicle de vida, els sistemes d'IA haurien de respectar els drets humans, la diversitat i l'autonomia de les persones. En aquest sentit, la pregunta guia és la següent: **Quin és l'impacte del vostre sistema d'IA envers l'autonomia de les persones?**

1. Impacte en les capacitats dels usuaris			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	Heu establert un procés d'avaluació del vostre sistema d'IA per garantir que en cap cas disminueix el nombre d'eleccions possibles, ni tampoc l'obliga o força a prendre unes determinades decisions?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
1.2	Heu establert un protocol de comprovació sobre el vostre sistema d'IA que permeti certificar que ajuda a les persones a prendre decisions millors i més informades d'acord amb els seus objectius?		ALTAI/OEIAC
1.3	Podeu assegurar que el vostre sistema d'IA, durant cap fase del seu cicle de vida, restringeix l'abast de les opcions d'estil de vida, creences, opinions, expressions o experiències personals?		UNESCO/ OEIAC
1.4	Heu establert els mecanismes per assegurar que tots els usuaris finals del sistema d'IA tenen prou capacitació i informació per ser-ne usuaris, incloent-hi les persones amb problemes d'accessibilitat o que es troben en risc d'exclusió?		UNESCO/ OEIAC

2. Percepció del sistema			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Heu establert mesures per tal que, en cap cas, les operacions del vostre sistema d'IA puguin veure's com a opaques, enganyoses o manipuladores per als usuaris?	Justificar resposta (500 caràcters max.)	ALTAI/OEIAC
2.2	Heu avaluat si el vostre sistema d'IA encoratja als seus usuaris a desenvolupar una inclinació, simpatia o empatia que pugui portar als usuaris a sobreestimar les capacitats del mateix sistema?		ALTAI/OEIAC

Sostenibilitat

Al llarg del seu cicle de vida, els sistemes d'IA haurien de beneficiar les persones, la societat i el medi ambient. En aquest sentit, la pregunta guia és la següent: **Tenint en compte els objectius de desenvolupament sostenible, els objectius del sistema d'IA es poden identificar i justificar clarament?**

1. Desenvolupament sostenible			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
1.1	Considereu que el vostre sistema d'IA té un impacte positiu en els drets laborals i/o en la qualitat del treball de les persones?	Justificar resposta (500 caràcters max.)	UNESCO/ OEIAC
1.2	Considereu que el vostre sistema d'IA millora la qualitat de vida dels seus usuaris a nivell individual i vetlla pel desenvolupament sostenible a nivell col·lectiu?		UNESCO/ OEIAC
1.3	Heu tingut en compte l'accés social, incloent-hi la bretxa digital, en l'ús i implantació del vostre sistema d'IA?		Attard-Frost et al. 2022/ OEIAC
1.4	Heu abordat els punts cecs de l'anàlisi mitjançant la participació amb les parts interessades rellevants, per exemple comprovant les hipòtesis i discutint les implicacions del vostre sistema d'IA amb les comunitats afectades?		UNESCO/ OEIAC

2. Impacte mediambiental			
Nº	**Indicador**	**Observable: Sí (1) /No (0)**	**Ref.**
2.1	Heu avaluat l'eficiència energètica del vostre sistema d'IA tenint en compte el seu cicle de vida?	Justificar resposta (500 caràcters max.)	ALTAI/ UNESCO/ OEIAC
2.2	Heu establert mesures per incrementar l'eficiència energètica del vostre sistema d'IA a través de la supervisió que, per naturalesa, és més eficient amb les dades?		ALTAI/ UNESCO/ OEIAC
2.3	Heu establert mesures per incrementar l'eficiència energètica del vostre sistema d'IA a través de la minimització de dades o de l'ús de dades sintètiques?		ALTAI/ UNESCO/ OEIAC
2.4	Heu establert mesures per incrementar l'eficiència energètica del vostre sistema d'IA transferint coneixement de dades riques a dominis de dades pobres?		ALTAI/ UNESCO/ OEIAC

3.5. Com realitzar l'autoavaluació a través de la web

El model PIO es troba disponible a través de l'adreça web www.oeiac.cat, des de la qual les diferents organitzacions i persones interessades poden accedir a diferents recursos com ara la informació general sobre el funcionament del model d'autoavaluació, l'informe amb la proposta general i finalment el qüestionari. En aquest sentit recomanem que aquest procés d'autoavaluació es realitzi en grups interdisciplinaris o en diferents fases per departaments ja que no és convenient que ho faci una sola persona.

A continuació detallem els passos més importants:

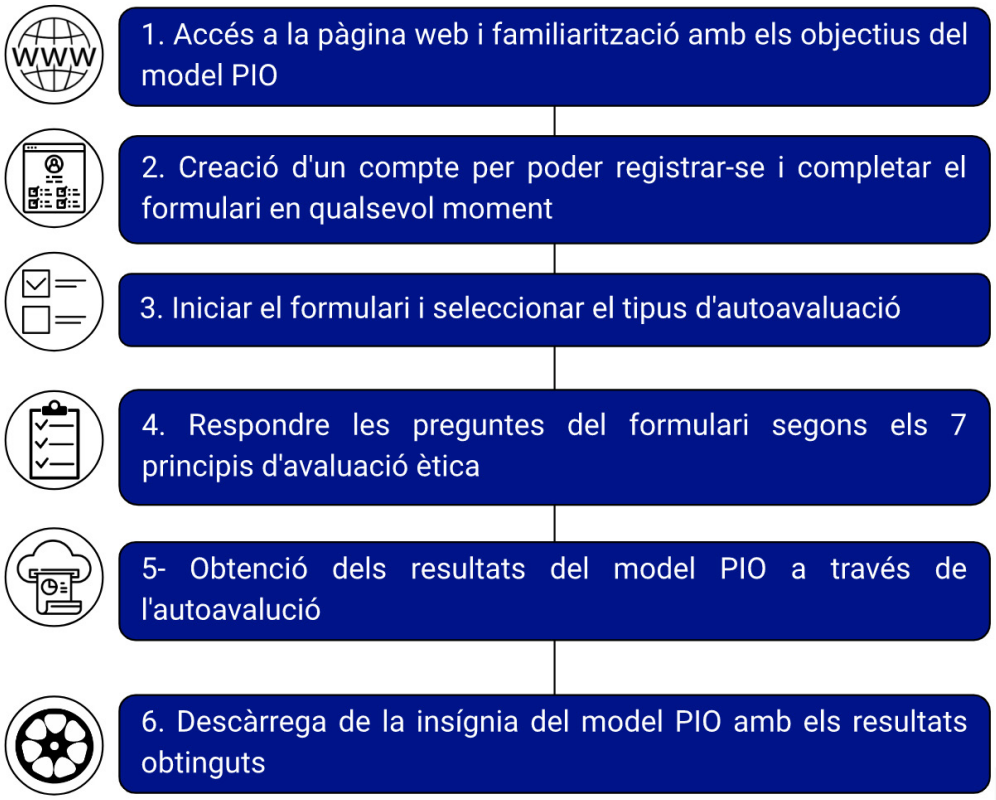


Fig. 5. Desenvolupament dels passos per portar a terme l'autoavaluació PIO i obtenir la insígnia

En la pàgina principal, es troba una presentació del propi model amb l'enllaç directe al formulari amb una simulació sobre la insígnia de qualitat que es genera amb els resultats del model d'autoavaluació.

Fig. 6. Pàgina principal oeiac.cat

En aquest mateix espai es pot trobar una breu explicació sobre la missió i els objectius del propi model PIO.

Fig. 7. Pàgina principal oeiac.cat

Des de la pàgina principal podem accedir al menú principal de navegació que està constituït pels següents apartats:

Qui som: En aquesta secció es pot trobar una breu descripció del OEIAC i les seves funcions, també es pot trobar informació sobre l'Estratègia d'Intel·ligència Artificial de la Generalitat de Catalunya CATALONIA.AI així com un enllaç directe al document oficial. En aquest mateix bloc d'informació es troba l'enllaç a la web oficial de l'OEIAC on es poden trobar totes les activitats i recursos descarregables realitzats fins avui.

Fig. 7. Presentació de la secció "Qui som" a oeiac.cat

Com funciona: En aquest apartat es pot trobar tota la informació en referent a la manera de funcionament del qüestionari del model PIO. Aquí es troben detallades la descripció de les fases i els dos tipus d'avaluació que els usuaris poden realitzar: l'autoavaluació ràpida i la completa. En l'última part es pot trobar informació sobre la insígnia de qualitat que és la imatge que es genera una vegada s'ha completat el formulari i que conté els resultats d'aquest.

En primer lloc, les organitzacions podran veure una presentació de cada principi i la selecció d'una sèrie d'indicadors pregunta seleccionats estratègicament després d'una revisió exhaustiva de la literatura especialitzada (Fase 1). En segon lloc, es recolliran totes les respostes afirmatives i negatives de les autoavaluacions ràpides i completes així com les justificacions a les respostes per aquestes últimes, si escau (Fase 2).

Amb la informació obtinguda de l'autoavaluació s'obtindran mètriques per cada un dels 7 principis avaluats i amb elles perfils observables de cada una de les organitzacions que portin a terme l'autoavaluació (Fase 3). Un cop obtingudes les mètriques corresponents, es procedirà a l'elaboració d'una insígnia o segell de qualitat que serà la imatge gràfica del model d'autoavaluació PIO representant l'adequació de les dades i del sistema d'IA segons els 7 principis ètics del model: (1) la transparència i explicabilitat, (2) la justícia i equitat, (3) la seguretat i no maleficència, (4) la responsabilitat i rendició de comptes, (5) la privacitat, (6) l'autonomia, i (7) la sostenibilitat. En aquest procés s'utilitza un gradient de color per representar d'una manera ràpida, senzilla i efectiva els resultats obinguts mitjançant el model d'autoavaluació (Fase 4).

En aquest sentit, abans de portar a terme l'autoavaluació serà important escollir la persona o persones responsables dins de les organitzacions públiques o privades que utilitzen sistemes d'IA per realitzar, des d'un punt de vista competencial, una autoavaluació amb la màxima informació disponible. Per fer-ho, es recolliran respostes binàries a 100 preguntes sobre cada un dels 7 principis anteriorment descrits, de manera que es recolliran dos valors possibles en les respostes (Si/No).

Es podrà realitzar una autoavaluació ràpida de posició i una autoavaluació de posició completa amb recomanacions. Aquesta última requereix la justificació de les respostes afirmatives, ja que això permetrà a l'equip d'OEIAC, juntament amb les persones col·laboradores externes, analitzar les respostes qualitatives i poder proporcionar feedback a l'organització que l'hagi realitzat si així ho desitja. Cal tenir en compte que, a partir dels principis, indicadors i observables, es pot proporcionar una descripció de l'orientació i adequació que pren un sistema d'IA no sols per a les persones que el desenvolupen, sinó també per als seus usuaris. D'aquesta manera, a més d'obtenir unes mètriques específiques i un segell de qualitat, les organitzacions que hagin realitzat l'autoavaluació completa rebran recomanacions específiques de millora.

Així doncs, a manera de resum aquestes són les fases que inclou cada model d'autoavaluació:
Autoavaluació ràpida: FASE 1+ FASE 2 + FASE 3 + FASE 4
Autoavaluació completa: FASE 1+ FASE 2 + FASE 3 + FASE 4 + Feedback OEIAC

Fig. 8. Presentació de la secció "Com funciona" a oeiac.cat

A qui va dirigit: En aquesta secció aprofundim en les raons per les quals realitzar el model d'autoavaluació PIO i descrivim qui pot ser el públic objectiu per a realitzar i obtenir aquesta certificació.

Fig. 9. Presentació de la secció "A qui va dirigit?" a oeiac.cat

Senzillesa: En aquest apartat aclarim alguns dubtes que puguin sorgir relacionats amb quina fase de desenvolupament del cicle de IA aplicar el model d'autoavaluació PIO.

Fig. 10. Presentació de la secció "Senzillesa" a oeiac.cat

Informe: En aquesta secció es troba disponible una còpia del mateix informe i es pot descarregar en diferents idiomes.

Fig. 11. Presentació de la secció "Informe" a oeiac.cat

Registre: per a dur a terme el formulari PIO és imprescindible donar-se d'alta com a usuari, aquest procés s'ha de fer en la secció de registre, encara que també es pot realitzar just abans de començar el qüestionari si no s'ha creat prèviament un usuari. Per al registre és necessari introduir el nom, un e-mail vàlid i una contrasenya robusta així com llegir i acceptar la política de protecció de dades i de privacitat.

Aquest registre permetrà a l'equip encarregat de dur a terme l'autoavaluació poder continuar-lo una vegada iniciat i reprendre el progrés en les següents sessions. Això és especialment rellevant si el qüestionari es realitza entre diversos departaments.

Tenint en compte que la seva realització requereix de capacitat de decisió, gestió i coneixement tècnics i socials del context d'implementació, és molt recomanable que l'autoavalució es faci en equip, per la qual cosa totes les persones que hi participen han de aportar informació identificativa.

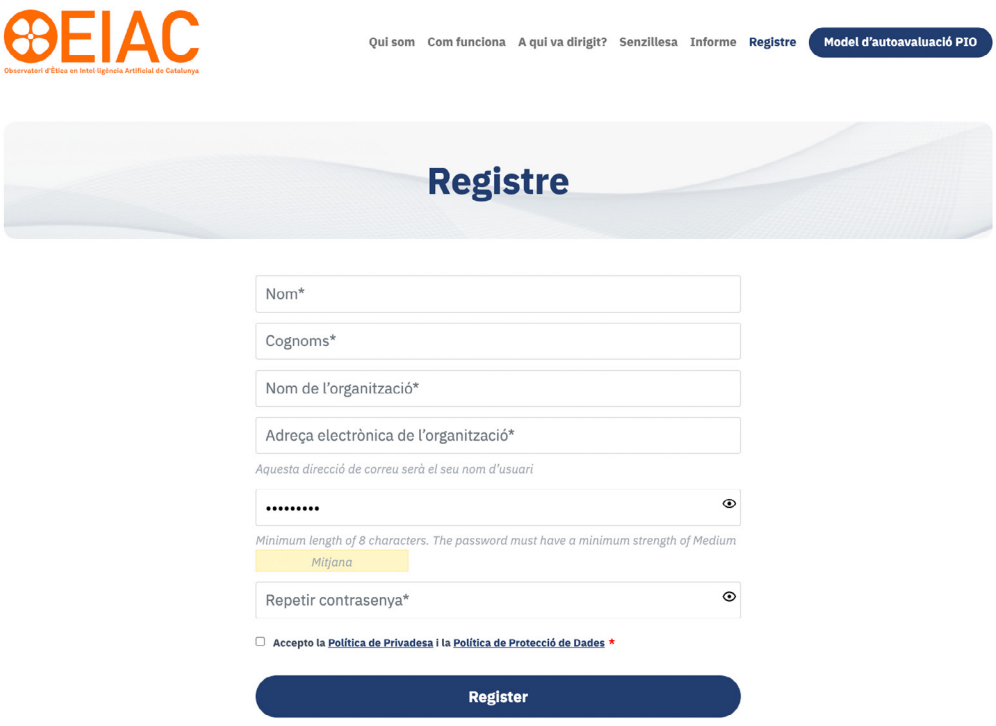


Fig. 11. Imatge del procés de registre a oeiac.cat

Donat que es recull informació personal així com de les organitzacions públiques i privades, en l'avaluació també s'inclou l'acceptació de la Política de Privades i la Política de Protecció de Dades.

Política de privadesa

Entenent que el dret a la protecció de dades és un dret fonamental, l'Observatori d'Ètica en Intel·ligència Artificial de Catalunya a la Universitat de Girona assumeix plenament els principis del Reglament (UE) 2016/679 del Parlament Europeu i del Consell, de 27 d'abril de 2016, relatiu a la protecció de les persones físiques pel que fa al tractament de dades personals i a la lliure circulació d'aquestes dades i pel qual es deroga la Directiva 95/46/CE (Reglament general de protecció de dades o RGPD, consultable des d'aquest enllaç) i de la Llei orgànica 3/2018, de 5 de desembre, de protecció de dades personals i garantia dels drets digitals (LOPDGDD, consultable des d'aquest enllaç), normes que donen garanties a les persones en relació al tractament de les seves dades.

Protecció de dades

L'Observatori d'Ètica en Intel·ligència Artificial de Catalunya (OEIAC), a través de la Universitat de Girona, es compromet a respectar la legislació vigent de protecció i confidencialitat en el tractament de dades de caràcter personal, d'acord amb el que disposa el Reglament (UE) núm. 2016/679 del Parlament Europeu i del Consell, de 27 d'abril de 2016, relatiu a la protecció de les persones físiques pel que fa al tractament de dades personals i a la lliure circulació d'aquestes dades.

El tractament de les dades personals que faciliteu queda legitimat pel consentiment explícit de l'interessat abans d'iniciar el qüestionari d'avaluació del model PIO. Les dades personals facilitades per mitjà del qüestionari s'utilitzaran amb la finalitat exclusiva de mantenir la relació administrativa derivada de la gestió de la sol·licitud d'informació, per enviar informació per correu electrònic o per activitats de l'OEIAC en relació a la proposta general d'avaluació de dades i sistemes d'intel·ligència artificial.

L'OEIAC es compromet a no comunicar les dades de caràcter personal subministrades per la persona interessada per cap mitjà i a no vendre, ni llogar, ni cedir les dades personals dels usuaris dels qüestionaris d'avaluació del model PIO, sense el consentiment del titular de les dades. Només les persones que hagin donat el consentiment explícit podran accedir, rectificar, oposar-se i suprimir les seves dades, així com exercitar el dret a la limitació del tractament de les dades. En tot cas, les dades personals només seran cedides als òrgans judicials i a autoritats administratives si una llei ho estableix preceptivament. El període de conservació de les dades serà el necessari per a la realització de l'avaluació del model PIO.

Les dades de caràcter personal seran tractades amb el grau de protecció legalment

requerit segons l'RGPD, i es prendran les mesures de seguretat necessàries per evitar-ne el mal ús, l'alteració, la pèrdua, el tractament o l'accés no autoritzat i la sostracció per part de tercers, que les poden utilitzar per a finalitats fraudulentes o diferents d'aquelles per a les quals han estat sol·licitades a l'usuari.

Les dades de navegació que recull el lloc web de l'OEIAC [www.oeiac.cat] no contenen informació que es pugui associar a usuaris concrets i només s'utilitzen per obtenir informació estadística sobre l'ús d'aquest lloc web. L'OEIAC no utilitza galetes (cookies), o altres mitjans de naturalesa anàloga, per tractar dades personals que permetin identificar persones físiques concretes, usuàries del lloc web. L'ús d'aquests mitjans està reservat a recollir informació tècnica que faciliti als usuaris l'accessibilitat i l'exploració segura i eficient d'aquest lloc, així com informació que permeti mesurar l'ús que se'n fa.

Si considereu que no hem donat una resposta adient, podeu adreçar-vos a la Universitat de Girona o podeu presentar una reclamació davant l'Autoritat Catalana de Protecció de Dades o exercir accions judicials.

Contacte:
Universitat de Girona
Edifici Àligues
Pl. Sant Domènec, 3
Campus Barri Vell
17004 – Girona

Iniciar formulari: Tant accedint des de la pàgina principal com a través del menú superior, iniciar formulari ens portarà directament al qüestionari d'autoavaluació PIO. El primer que trobarem en aquesta secció serà una breu introducció al model, instruccions sobre com emplenar-lo i una clarificació sobre els dos tipus d'autoavaluació que es poden realitzar, la ràpida i la completa

Fig. 12. Imatge de la presentació del Model d'autavaluació PIO a oeiac.cat

En fer clic en el botó "Iniciar formulari" ens trobarem amb la pàgina de login en la qual hem d'introduir el nostre compte. En cas que no estiguem registrats, haurem de realitzar tot el procés.

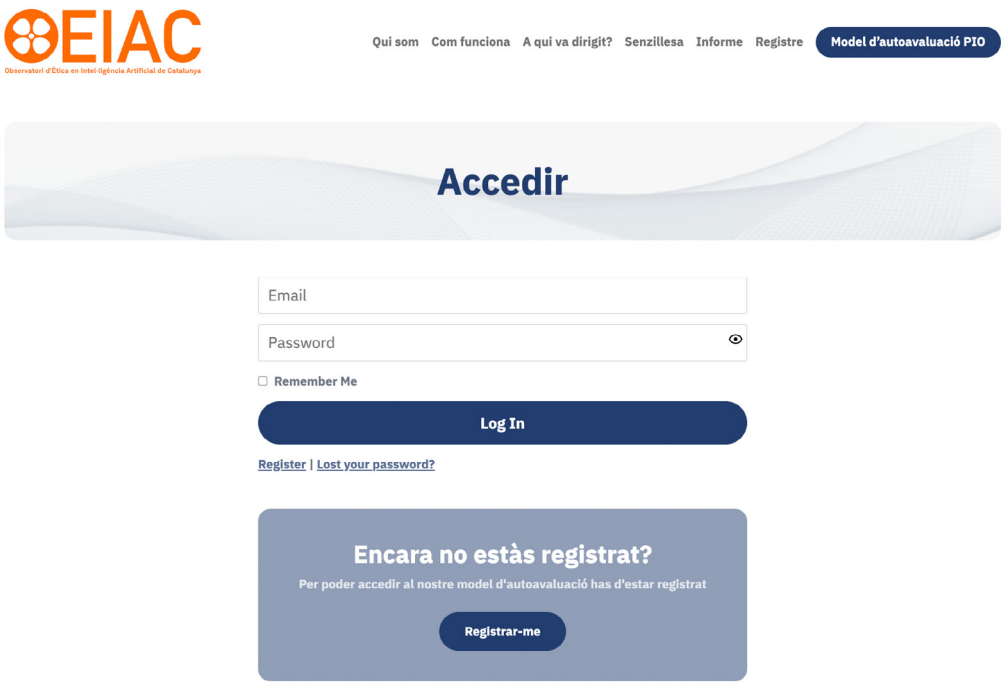


Fig. 13. Imatge de la pàgina per accedir al Model d'autoavaluació PIO

Seguidament haurem de triar el tipus d'avaluació que volem dur a terme. La diferència entre aquests dos tipus d'avaluació fan referència a la quantitat d'informació que haurem d'aportar en cada pregunta així com al feedback final que rebrem. En l'autoavaluació ràpida

els responsables que estiguin emplenant el formulari hauran de respondre a cadascuna de les preguntes observables dels principis ètics amb "si", "no" i "no aplica", i al final del procés rebran la insígnia de qualitat amb les respostes. D'altra banda, aquells responsables que decideixin realitzar l'autoavaluació completa, a més de respondre a cadascuna de les preguntes observables hauran de justificar la seva resposta, això els permetrà obtenir recomanacions per part de l'equip de l'OEIAC si ho sol·liciten.

Encara que hem d'assenyalar que la decisió sobre realitzar l'un o l'altre tipus d'avaluació es pot modificar en qualsevol part del procés. Així mateix, és important esmentar que en aquesta pantalla haurem de llegir i acceptar la declaració de responsabilitat mitjançant la qual s'estipula que els usuaris es fan responsables que les respostes que donin siguin verídiques. Més concretament, les organitzacions o persones hauran d'estar conformes amb la següent declaració:

Declaració responsable

En qualitat de persona o persones que avaluen l'organització, declaro, sota la meva responsabilitat, que proporciono al present qüestionari d'avaluació del model PIO, informació i documentació veraç, vigent, exacte i complerta. La inexactitud, la falsedat o l'omissió de caràcter essencial en qualsevol informació, manifestació o document que formi part, determinarà, prèvia la instrucció del procediment oportú, la denegació de l'avaluació corresponent i la resolució de la mateixa si aquesta ja hagués estat atorgada. Tot això, sens perjudici de les responsabilitats penals, civils o administratives en que hagi pogut incórrer la persona o persones que avaluen l'organització.

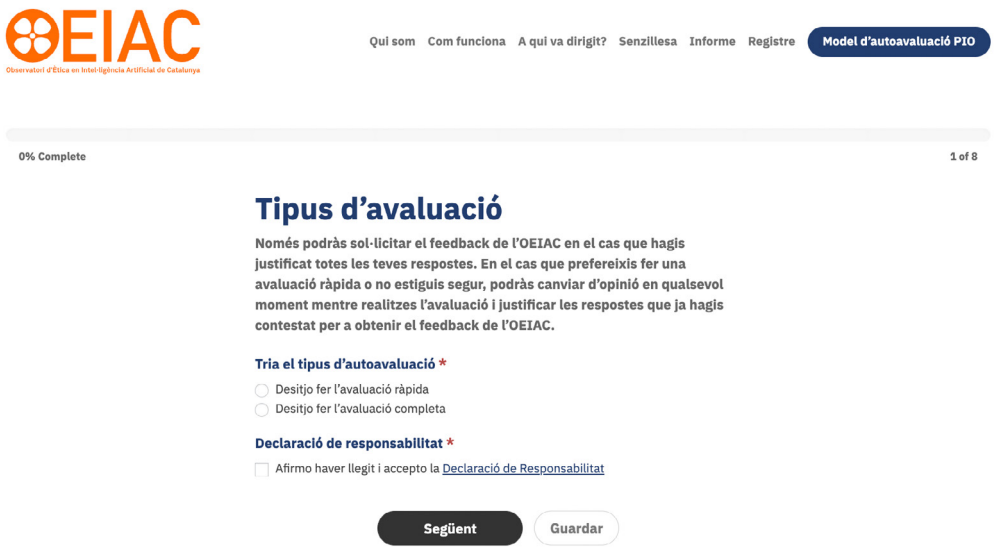


Fig. 15. Imatge de la pàgina per seleccionar el tipus d'avaluació

Una vegada fem clic en l'opció de Següent, accedirem directament al qüestionari i a la primera categoria de preguntes que són les relacionades amb el principi de Transparència i explicabilitat. En cada principi trobarem una sèrie d'indicadors (en l'exemple podem veure "1. Anàlisi preliminar de l'ús del sistema") i unes preguntes associades a aquests indicadors. Cada pregunta s'ha de respondre fent clic en "sí", "no" o "no aplica" depenent del cas. És important assenyalar que la secció de "Justifica la teva resposta" està només disponible en el cas que s'estigui realitzant una autoavaluació completa.

En la part superior podem trobar una barra amb el nostre progrés i en la part inferior de la pantalla podem guardar el formulari fent clic en l'opció Save draft, el que ens permet poder recuperar-lo més endavant si així ho considerem.

Fig. 16.Imatge del qüestionari relacionat amb el principi de Transparència i explicabilitat

Una vegada completades les 70 preguntes observables dels principis ètics proposats en el model PIO, ens apareixerà el resultat amb les nostres respostes i el percentatge de preguntes afirmatives i negatives associades a cada principi. Aquest resultat està supeditat a la pròpia declaració de responsabilitat, mitjançant la qual, l'organització es fa responsable de la veracitat de les respostes aportades.

En aquesta secció també ens apareixerà la insígnia amb els resultats i podrem descarregar la imatge en format png per a utilitzar-la en altres plataformes. És important assenyalar que aquesta insígnia no és un element certificador com a tal sinó una eina d'autoavaluació que reflecteix la situació d'una organització d'acord amb els 7 principis definits en el model.

Al final de aquesta pantalla podràs posar-te en contacte amb l'equip de l'OEIAC per a interpretar els resultats, obtenir recomanacions i resoldre els dubtes pendents.

Fig. 17. Imatge dels resultats obtinguts al final del qüestionari i insígnia PIO

Finalment, després de repassar totes les seccions que comprenen la pàgina web i a manera de resum, destaquem els principals passos que s'han de donar per a realitzar el procés d'autoavaluació ètica PIO:

1- Accés a la pàgina web i familiarització amb els objectius proposats.
2- Crear un compte i registrar-se.
3- Iniciar el formulari i seleccionar el tipus d'autoavaluació.
4- Respondre a les preguntes del formulari atesos els 7 principis ètics.
5- Obtenir els resultats del model PIO.
6- Descarregar la insígnia PIO amb els valors resultants.

Agraïments

Volem agrair tots els comentaris rebuts des del principi en la concepció, desenvolupament i realització d'aquest treball als membres del Consell Assessor de l'Observatori d'Ètica en Intel.ligència Artificial de Catalunya: Artur Serra, Carina Lopes, Carme Torras, David Pereira, Elisabet Golobardes, Fernando Vilariño, Itziar de Lecuona, Joan Mas, Karina Gibert, Karma Peiró, Liliana Arroyo, Marc Pérez-Batlle, Miquel Domènech, Montse Guàrdia, Nuria Oliver, Ramon López de Mántaras, Ramon Trias, Ricardo Baeza-Yates, Ulises Cortés i Xavier Trabado.

També volem agrair els suggeriments rebuts de Daniel Marco (Director General d'Innovació i Economia Digital, Generalitat de Catalunya), Josep Calbó (Vicerector de Projectes Estratègics de la Universitat de Girona), Daniel Santanach (Coordinador de l'Estratègia de IA de Catalunya (CATALONIA.AI) al Departament de la Vicepresidència i de Polítiques Digitals i Territori de la Generalitat de Catalunya) i Àngel Martín Carballo (Fundació i2CAT).

Finalment, volem donar les gràcies a les persones que van participar en els seminaris d'introducció i capacitació tècnica "Principis, Indicadors i Observables (PIO): Una proposta general d'avaluació ética de dades i sistemes d'intel·ligència artificial" els dies 27 i 28 d'abril de 2022 per ajudar a millorar aquest treball: Ana Belén Elies, Anna Quer i Rodríguez, Àngel Martín, Ángeles Bueno, Àngels Vidal, Alba Latorre, Artur Llobera, Aysha Ait, Carina Lopes, Carles Pla, Carolina Moltó, Cristian Barrué, Elena Bilbao, Eloi Mayordomo, Ernest Pons Fanals, Fernanda de la Torre, Francesc Quintana, Hector Garcia Morales, Irene Rodenas, Joan Segur, Josep Maria Flores, Juan Gonzalez Fraile, Lluís Sabaté, Lluïsa Parés, Lourdes Petit, Mar Sibila, Marc Pérez-Batllé, Marta Rodríguez, Miguel Angel de Manuel, Miguel Sirvent, Miriam Carles Soler, Núria Abdón, Patricia Roca, Pau Alcantara, Paula Boet, Ramon Trias, Sarah Dubois, Susanna Aussó i Tanya Alvarez.

Bibliografia

ACM (2017) Statement on algorithmic transparency and accountability, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf, (consultada l'11 de març de 2022)

Ada Lovelace Institute. (2021, diciembre). Technical methods for regulatory inspection of algorithmic systems in social media platforms. https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/ (consultada el 10 de març de 2022)

Ada Lovelace Institute, AI Now Institute & Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/ (consultada el 25 de febrer de 2022).

Ada Lovelace Institute & DataKind, (2020). Examining the black box: Tools for assessing algorithmic systems. Technical report, AdaLovelace Institute, https://ico. org. uk/media/about-theico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation. pdf. (consultada el 8 d'abril de 2022)

AI HEG (2019) Ethics guidelines for trustworthy AI. European Ethics guidelines for trustworthy AI, Publications Office, https://data.europa.eu/doi/10.2759/177365 (consultada l'11 de març de 2022)

AI HLEG (2020) The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, European Commission, Brussels, https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence (consultada l'11 de març de 2022)

AI Now Institute (2018, octubre) Algorithmic Accountability Policy Toolkit (N.o 01), https://ainowinstitute.org/aap-toolkit.pdf

AIEI Group (2020) AI Ethics Impact Group: From Principles to Practice – An interdisciplinary framework to operationalise AI ethics, https://www.ai-ethics-impact.org/en (consultada el 27 de agosto de 2021).

Algo.Rules (2019) "Rules for the design of algorithmic systems", https://algorules.org/en/home (consultada l'11 de març de 2022)

Allistene (2014) Ethique de la recherche en Robotique, CERNA, http://cerna-ethics-allistene.org/digitalAssets/38/38704_Avis_robotique_livret.pdf

Anderson, D., Bonaguro, J., McKinney, M., Nicklin, A., & Wiseman, J. (2018). Ethics & Algorithms Toolkit (beta). Ethics Toolkit, de https://ethicstoolkit.ai/ consultada el 23 de juny de 2022).

Attard-Frost, B., De los Ríos, A., & R Walters, D. (2022). The Ethics of AI Business Practices: A Review of 47 AI Ethics Guidelines,https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4034804

Australian Government: Department of industry, Science, Energy and Resources. (2019). Australia's Artificial Intelligence Ethics Framework, https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework (consultada el 23 de juny de 2022).

Barcelona Declaration (2017) Barcelona Declaration for the Proper Development and Usage of Artificial Intelligence in Europe, IIIA CSIC, https://www.iiia.csic.es/barcelonadeclaration/ (consultada el 27 de agosto de 2021).

Campolo, A., Sanfilipo, M., Whittaker, M., & Crawford, K. (2017, octubre). AI Now 2017 Report (A. Selbst & S. Barocas, Eds.), https://ainowinstitute.org/AI_Now_2017_Report.pdf, (consultada l'11 de març de 2022)

CATALONIA. AI (2020) L'Estratègia d'Intel·ligència Artificial de Catalunya, Generalitat de Catalunya, Departament de Vicepresidència i de Polítiques Digitals i Territori, https://politiquesdigitals.gencat.cat/ca/tic/catalonia-ai (consultada el 9 de març de 2022)

Central Digital and Data Office. (2020, 16 de setembre). Data Ethics Framework. United Kingdom Government, https://www.gov.uk/government/publications/data-ethics-framework (consultada el 23 de juny de 2022).

Central Digital and Data Office. (2022, 1 de juny). Algorithmic Transparency Standard. United Kingdom Government, https://www.gov.uk/government/collections/algorithmic-transparency-standard (consultada el 23 de juny de 2022).

Center for Data Science and Public Policy, University of Chicago. (2018). Bias and Fairness Audit Toolkit. Aequitas: Bias & Fairness Audit, http://aequitas.dssg.io/(consultada el 23 de juny de 2022).

Clark, R. (2019) Principles for AI: a SourceBook, http://www.rogerclarke.com/EC/GAIP-

190210.html (consultada l'11 de març de 2022)

Collectif, C. (2018). Research ethics in machine learning (CERNA; ALLISTENE), https://hal.archives-ouvertes.fr/hal-01724307/ (consultada l'11 de març de 2022)

COMEST/UNESCO (2017), Report of COMEST on robotics ethics, https://unescoblob.blob.core.windows.net/pdf/UploadCKEditor/REPORT%20OF%20COMEST%20ON%20ROBOTICS%20ETHICS%2014.09.17.pdf

Datatilsynet (2018) Artificial intelligence and privacy, https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf (consultada l'11 de març de 2022)

Demiaux, V., & Si Abdallah, Y. (2017). How can humans keep the upper hand. Report on the Ethical Matters Raised by Algorithms and Artificial Intelligence. Paris: Commision Nationale Informatique et Libertés.

Deon (s. f.). An ethics checklist for data scientists, https://deon.drivendata.org/ (consultada el 23 de juny de 2022).

Digital Dubai Office (2020) AI System Ethics Self-Assesment Tool, https://www.digitaldubai.ae/self-assessment (consultada el 23 de juny de 2022).

European Commission (2019, 8 d'abril). Ethics Guidelines for Trustworthy AI – Independent High-Level Expert Group on Artificial Intelligence, https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence
(consultada l'11 de març de 2022)

European Commission (2020). White Paper on Artificial Intelligence: A European approach to excellence and trust (COM(2020) 65). Brussels: European Commission, https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (consultada el 27 d'agost de 2021).

European Commission (2021). Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Brussels: European Commission, https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (consultada el 25 de febrer de 2022).

European Group on Ethics in Science and New Technologies (EGE) (2018). Statement on Artifcial Intelligence, Robotics and 'Autonomous' Systems, https://data.europa.eu/doi/10.2777/786515 en (consultada el 8 de març de 2022).

European Law Institute. (2022). Guiding Principles for Automated Decision-Making in the EU. ELI Innovation Paper, https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Innovation_Paper_on_Guiding_Principles_for_ADM_in_the_EU.pdf (consultada el 17 de maig de 2022).

European Parliament (2017), Report with recommendations to the commission on civil law rules on robotics, https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.html (consultada l'11 de març de 2022)

Fairnlearn (2021, 7 juliol) FairLearn. GitHub, https://github.com/fairlearn/fairlearn (consultada el 23 de juny de 2022).

FATML (2016) Principles for accountable algorithms and a social impact statement for algorithms, https://www.fatml.org/resources/principles-for-accountable-algorithms (consultada l'11 de març de 2022)

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V. & Vayena, E. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds and Machines, 28(4): 689-707.

Gilburt, B. (2019) Women leading in AI: 10 principles of responsible AI, Towards Data Science, https://towardsdatascience.com/women-leading-in-ai-10-principles-for-responsibleai-8a167fc09b7d (consultada l'11 de març de 2022)

Green Digital Working Group (2016) Position on robotics and artificial intelligence, https://felixreda.eu/wp-content/uploads/2017/02/Green-Digital-Working-Group-Position-on-Robotics-and-Artificial-Intelligence-2016-11-22.pdf (consultada l'11 de març de 2022)

Government of Canada. (2022, 25 de febrer). Algorithmic Impact Assessment, de https://open.canada.ca/aia-eia-js/?lang=en (consultada el 23 de juny de 2022).

IBM Research (2019). AI Explainability 360. IBM Research Trusted AI, http://aix360.mybluemix.net/(consultada el 23 de juny de 2022).

ICO (2017), Big data, artificial intelligence, machine learning and data protection, https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.

pdf (consultada l'11 de març de 2022)

ICO (2021). AI and Data Protection Risk Mitigation And Management Toolkit. https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-artificial-intelligence-and-data-protection/ (consultada l'11 de març de 2022)

IEEE (2019). Ethically Aligned Design - A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, https://standards.ieee.org/content/dam/ieeestandards/standards/web/documents/other/ead_v2.pdf (consultada el 25 de febrer de 2022).

IEEE (2021) IEEE CertifAIEd, https://engagestandards.ieee.org/ieeecertifaied.html (consultada el 23 de juny de 2022)

IEEE (2021) The IEEE CertifAIEd Framework for AI Ethics Applied to the City of Vienna. IEEE SA, https://beyondstandards.ieee.org/the-ieee-certifaied-framework-for-ai-ethics-applied-to-the-city-of-vienna/ (consultada el 8 de març de 2022)

IEEE (2022). IEEE Standard for Transparency of Autonomous Systems. IEEE SA, https://standards.ieee.org/ieee/7001/6929/?utm_source=beyondstandards&utm_medium=post&utm_campaign=working-group-2022&utm_content=7001 (consultada el 13 de maig de 2022

International Conference of Data Protection and Privacy Commissioners (ICDPPC) (2018), Declaration on ethics and data protection in artificial intelligence, https://icdppc.org/wpcontent/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf (consultada l'11 de març de 2022)

Internet Society (2017) Artificial intelligence and machine learning: policy paper, https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/ (consultada l'11 de març de 2022)

Institute For The Future (2020).Ethical OS Toolkit, https://ethicalos.org/(consultada el 23 de juny de 2022).

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

Latonero, M. (2018), "Governing artificial intelligence: upholding human rights and dignity", Data and Society.

Lernende Systeme. (s. f.). Lernende Systeme: Publications, https://www.plattform-lernende-systeme.de/publications.html (consultada el 23 de juny de 2022)

Leslie, D. (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.324052920

Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020, abril). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Madiega, T. (2021)Artificial intelligence act, EPRS: European Parliamentary Research Service, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf (consultada el 23 de juny de 2022).

Microsoft (2021, 23 setembre). InterpretML. GitHub, https://github.com/interpretml/interpret (consultada el 23 de juny de 2022)

Microsoft (2022) AI Fairness Checklist. Microsoft Research, https://www.microsoft.com/en-us/research/project/ai-fairness-checklist/ (consultada el 10 de març de 2022)

Microsoft (2022, 6 maig). Responsible Innovation toolkit: Azure Application Architecture Guide. Microsoft Docs, https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/ (consultada el 23 de juny de 2022).

Mittelstadt, B. (2016). Auditing for transparency in content personalization systems. International. Journal of Communication, 10 (Juny), 4991–5002.

Mittelstadt, B. (2019), "Principles alone cannot guarantee ethical AI", Nature Machine Intelligence, https://doi.org/doi:10.1038/s42256-019-0114-4 (consultada l'11 de març de 2022)

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2019), "From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices", Science and Engineering Ethics, available at: https://doi.org/10.1007/s11948-019-
00165-5

NSTC (2016) The national artificial intelligence research and development strategic plan, EEUU, https://www.nitrd.gov/pubs/national_ai_rd_strategic_plan.pdf (consultada l'11 de

març de 2022)

O'Neil, C. (2016), Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown Publishers.

OECD (2019) Recommendation of the Council on Artificial Intelligence, Committee on Digital Economy Policy, OECD Legal Instruments, https://legalinstruments.oecd.org/en/instruments/OECDLEGAL-0449#committees (consultada el 25 de febrer de 2022)

OECD (2022) OECD Framework for the Classification of AI systems", OECD Digital Economy Papers, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en. (consultada l'11 de març de 2022)

Personal Data Protection Comission Singapore (PDPC) (2022, 25 de maig) Singapore's Approach to AI Governance, https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework (consultada el 23 de juny de 2022).

Pymetrics (2020, 29 juliol). AI Audit. GitHub, https://github.com/pymetrics/audit-ai (consultada el 23 de juny de 2022).

PWC (2019) Responsible AI Toolkit, https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html (consultada el 23 de juny de 2022).

Rathenau Institute (2017) Human rights in the robot age: Challenges arising from the use of robotics, artificial intelligence, and virtual and augmented reality, https://www.rathenau.nl/en/digitale-samenleving/human-rights-robot-age (consultada l'11 de març de 2022).

Rosales Torres, C. S., Buenadicha Sánchez, C., & Narita, T. (2021, maig) Autoevaluación ética de IA para actores del ecosistema emprendedor: Guía de aplicación. Banco Interamericano de Desarrollo, http://dx.doi.org/10.18235/0003269

Sage (2017), "The ethics of code: Developing AI for business with five core principles". https://www.sage.com/~/media/group/files/business-builders/business-builders-ethics-of-code.pdf

SAP (2018) SAP's guiding principles for artificial intelligence (AI), https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/ (consultada l'11 de març de 2022)

SIIA (2017) Ethical principles for artificial intelligence and data analytics, https://www.

pr.com/press-release/735528 (consultada l'11 de març de 2022)

Special Interest Group on Artificial Intelligence (2018), "Dutch artificial intelligence manifesto, http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf (consultada l'11 de març de 2022)

The Public Voice (2018) Universal guidelines for artificial intelligence, https://thepublicvoice.org/ai-universal-guidelines/ (consultada l'11 de març de 2022)

TIETO (2019, febrer). ANNUAL REPORT 2018: Towards a sustainable data-driven world. https://www.tietoevry.com/siteassets/files/investor-relations/2018/tieto-annual-report-2018.pdf (consultada l'11 de març de 2022)

UNESCO (2021). Preliminary report on the first draft of the recommendation on the ethics of artificial intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000374266 (consultada l'11 de març de 2022)

UNI Global Union (2017) Top 10 principles for ethical AI, http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf (consultada l'11 de març de 2022)

United Nations Development Group (UNDG) (2017), "Data privacy, ethics and protection. Guidance note on big data for achievement of the 2030 agenda", https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf (consultada l'11 de març de 2022)

Uttamchandani, S. (2021, 30 d'agost). The AI Checklist from Idea to Production | Towards Data Science. Medium, https://towardsdatascience.com/the-ai-checklist-fe2d76907673 (consultada el 10 de març de 2022)

Vallor, S., Green, B. & Raicu , I. (2018) Ethics in Technology Practice. The Markkula Center for Applied Ethics at Santa Clara University, https://www.scu.edu/ethics/.(consultada el 23 de juny de 2022).

Vinuesa, R., Azizpour, H., Leite, I. et al. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. Nat Commun 11, 233. https://doi.org/10.1038/s41467-019-14108-y

Wagner, B. (2018). Ethics as an escape from regulation. From "ethics-washing" to ethics shopping? Being Profiled (pp. 84-89). Amsterdam: Amsterdam University Press.

Weinberger, D. (2018). Playing with AI Fairness. What-If Tool, https://pair-code.github.io/what-if-tool/ai-fairness.html (consultada el 23 de juny de 2022).

Wiener, N. (1954). The Human Use of Human Beings: Cybernetics and Society, Houghton Mifflin; 2nd edn Doubleday Anchor.

World Economic Forum (2018), White paper: How to prevent discriminatory outcomes in machine learning, https://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf (consultada l'11 de març de 2022)

World Economic Forum (2020) Empowering AI Leadership, https://www.weforum.org/projects/ai-board-leadership-toolkit (consultada el 23 de juny de 2022).

World Economic Forum (2021, diciembre) Human-Centred Artificial Intelligence for Human Resources: A Toolkit for Human Resources Professionals. https://www3.weforum.org/docs/WEF_Human_Centred_Artificial_Intelligence_for_Human_Resources_2021.pdf (consultada l'11 de març de 2022)