

Bazele algoritmice și etica în IA: de la teorie la practică

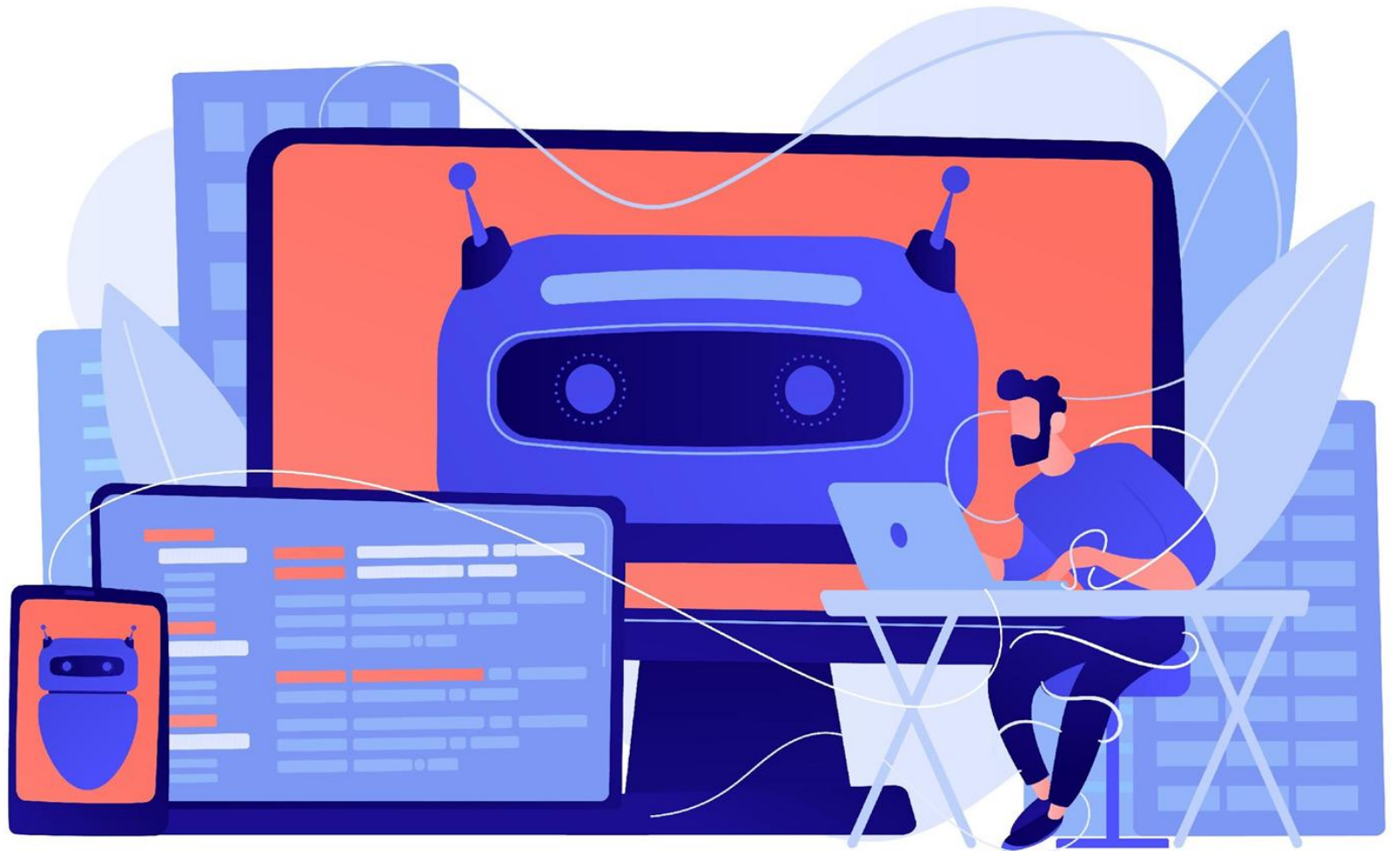
Set de instrumente pentru sesiuni
sincrone

CU1 | Etica IA - o abordare practică
Suport pentru diapozitive PowerPoint

INDEX

- INTRODUCERE - 3
- PRINCIPII ETICE ALE IA - 8
- CADRE ETICE, ORIENTĂRI ȘI SETURI DE INSTRUMENTE ALE IA - 23
- EU AI ACT - 33
- CICLUL DE VIAȚĂ AL DEZVOLTĂRII AI - 40
- IMPLICAREA PĂRȚILOR INTERESATE ÎN DEZVOLTAREA SOLUȚIILOR IA - 45
- PRINCIPII ETICE ȘI IMPLICAREA PĂRȚILOR INTERESATE ÎN CICLUL DE VIAȚĂ AL DEZVOLTĂRII - 53

INTRODUCERE



Sursa imaginii | Freepik

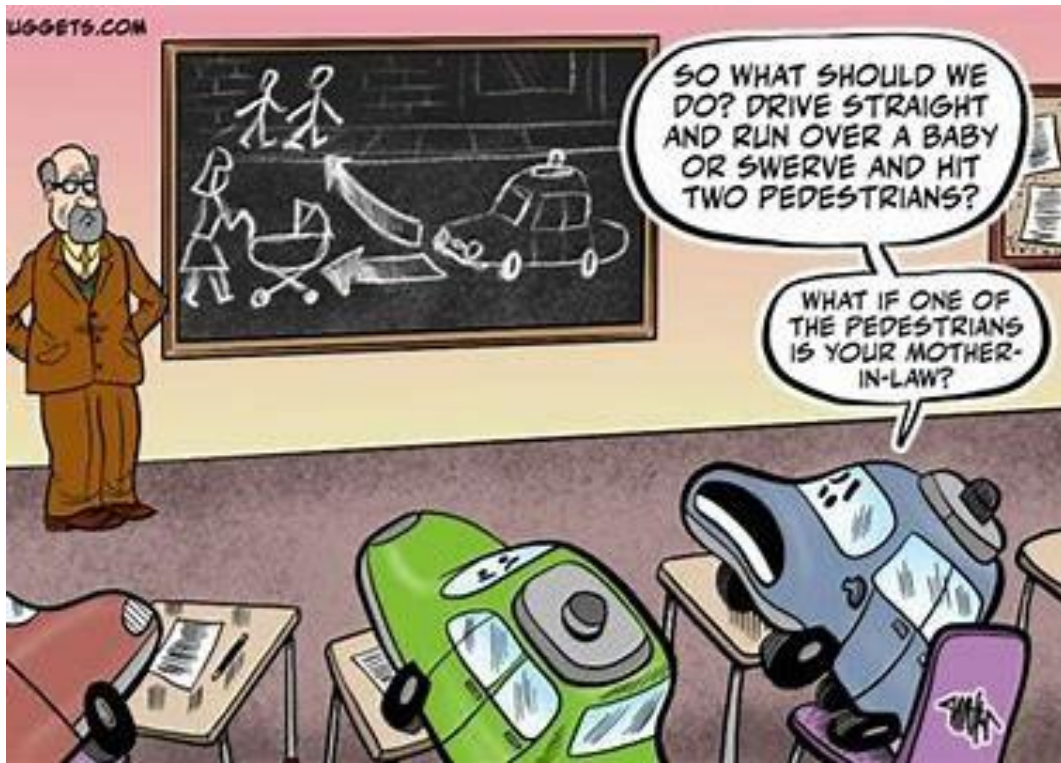
ÎN ACEASTĂ UNITATE DE COMPETENȚĂ VEȚI GĂSI URMĂTOARELE SUBIECTE:

- Etica IA și importanța acesteia în dezvoltarea soluțiilor IA
- Principiile etice ale IA
- Cadre etice, orientări și seturi de instrumente
- Grup de experți la nivel înalt; AI demn de încredere - cadru
- Actul UE privind inteligența artificială
- Ciclul de viață al dezvoltării AI
- Implicarea părților interesate în ciclul de viață al dezvoltării IA
- Ciclul de viață al dezvoltării IA cu principiile etice și părțile interesate aferente

LA SFÂRȘITUL UNITĂȚII DE COMPETENȚĂ, AR TREBUI SĂ FIȚI CAPABIL SĂ:

- Să fie conștienți de principiile etice în dezvoltarea și implementarea IA
- Recunoașterea diverselor cadre și orientări etice și aplicarea lor în scenarii din lumea reală care implică sisteme AI
- să înțeleagă importanța principiilor etice și a implicării părților interesate în procesul de dezvoltare etică a IA

Preocupări etice - Ce se întâmplă dacă...



Sursa imaginii | [Desen animat: Predarea eticii către AI - KDnuggets](#)

- AI ar avea puterea de a decide dacă primiți sau nu un împrumut?
- AI ar judeca cine va fi condamnat la închisoare?
- O mașină autonomă ar trebui să decidă dacă să salveze viața pietonului sau viețile celor din mașină?
- Ce s-ar întâmpla dacă un sistem de recrutare bazat pe inteligență artificială ar decide dacă să angajeze o femeie sau un bărbat ca și căpitan de zbor?

Inteligența artificială revoluționează viața



Sursa imaginii | OCDE; IA: Risc și oportunitate

Video de la: <https://youtu.be/-CXkHs3cxa4>

- Inteligența artificială are impact asupra tuturor factorilor vieții, influențând munca, timpul liber și oferind soluții la probleme globale precum schimbările climatice și accesul la asistență medicală, punând în același timp provocări semnificative pentru guverne și persoane.
- Integrarea inteligenței artificiale în economii și societăți ridică întrebări cu privire la cadrele politice și instituționale adecvate necesare pentru a ghida dezvoltarea și aplicarea acesteia în beneficiul societății.
- Considerațiile etice generale au condus la crearea și publicarea a numeroase abordări pentru a asigura aplicarea etică a IA, care sunt abordate în acest curs.

PRINCIPII ETICE ALE AI



Principiile etice ale IA

Care sunt principiile etice ale IA?

- **Principiile etice sunt definite ca orientări generale care abordează preocupările etice ale sistemelor AI.** Aceste principii sunt adesea abstracte și nu oferă instrucțiuni specifice cu privire la modul de aplicare a acestora într-un anumit context.
- Principiile etice descriu ceea ce se așteaptă în termeni de bine și rău și alte standarde etice. Principiile etice ale IA se referă la constrângerile normative cu privire la ceea ce **trebuie și ceea ce nu trebuie să** se facă în cazul utilizării algoritmilor în societate.

(Prem 2023), (Zhou et al., 2020)

Principiile etice ale IA

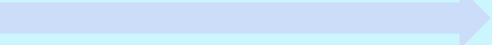
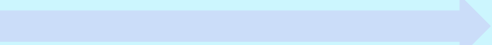
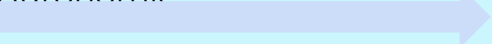
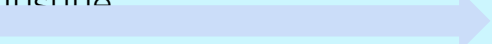
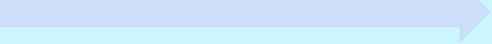
- O multitudine de recomandări au fost publicate pentru a ghida utilizarea etică a inteligenței artificiale.
- Contribuțiile provin din diverse sectoare, de ex:
 - Recomandarea OCDE, schița principiilor eticii IA (Publicat în mai 2019)
 - Principiile Beijing AI pentru cercetare și dezvoltare (Publicat în mai 2019)
 - Industrie (de exemplu, Google, IBM, Microsoft, Intel)
 - Guvern (de exemplu, Declarația de la Montreal, Grupul de experți al Comisiei Europene)
 - Mediul academic (de exemplu, Future of Life Institute, IEEE, AI4People)

(Prem 2023), (Morley et al 2020)

Principiile etice ale IA

Exemple de principii etice

Graficul arată modul în care diferite cadre combină principiile etice și modul în care acestea se suprapun. Săgețile orizontale leagă subiectele între ele.

AI4People (publicat în noiembrie 2018) (Floridi et al. 2018)	Cinci principii esențiale pentru orice cadru etic pentru IA (L Floridi și Clement-Jones 2019)	Orientări etice pentru inteligența artificială demnă de încredere (publicate în aprilie 2019) (Comisia Europeană 2019)	Recomandarea Consiliului pentru inteligență artificială (publicată în mai 2019) (OCDE 2019b)
Beneficență 	IA trebuie să fie benefică pentru omenire	Respectul pentru autonomia umană	Creștere favorabilă incluziunii, dezvoltare durabilă și bunăstare
Non-maleficență 	IA nu trebuie să încalce dreptul la viață privată sau să submineze securitatea	Prevenirea daunelor	Robustețe, securitate și siguranță
Autonomie 	IA trebuie să ne protejeze și să ne consolideze autonomia și capacitatea de a lua decizii și de a alege între alternative		Valori centrate pe om
Justiție 	IA trebuie să promoveze prosperitatea	Echitate	Echitate
Explicabilitate 	Sistemele AI trebuie să fie inteligibile și explicabile	Explicabilitate	Transparență și explicabilitate Responsabilitate

Modificat după Morley et al 2020



Sursa imaginii | Imagine dezvoltată de VAMK

Principiile etice ale IA prezintă o suprapunere semnificativă, sugerând o convergență către un set unificat de orientări etice.

- Benefic și respectuos față de oameni și mediu (**beneficență**).
- Robust și sigur (**non-maleficență**).
- Respectă valorile umane (**autonomia**).
- Corect (**justiție**).
- Explicabil, responsabil și inteligibil (**explicabilitate**).

Morley et al (2020), Jobin et al (2019)

Principiile etice ale IA



Sursa imaginii | Freepik

Beneficență

Beneficența în IA cuprinde promovarea bunăstării umanității

Principiul beneficenței pune accentul pe crearea IA în beneficiul oamenilor și al planetei, subliniind importanța promovării bunăstării și a durabilității mediului.

(Floridi et al 2018)

Principiile etice ale IA



Sursa imaginii | Freepik

Exemplu privind principiul Beneficenței

Recomandări privind asistența medicală

- Să luăm în considerare un sistem de asistență medicală bazat pe inteligență artificială care recomandă planuri de tratament pe baza datelor pacientului. Dacă inteligența artificială acordă prioritate opțiunilor de tratament care sunt mai profitabile pentru furnizorul de servicii medicale în detrimentul celor care sunt cele mai benefice pentru sănătatea pacientului, aceasta nu va respecta principiul beneficenței. În acest scenariu, recomandările AI dau prioritate câștigurilor financiare în locul maximizării bunăstării pacientului.
- Watson for Oncology de la IBM s-a confruntat cu provocări etice, deoarece a recomandat tratamente care nu au fost întotdeauna în concordanță cu orientările medicale sau cu interesele pacienților, ridicând probleme legate de prioritizarea profitului în detrimentul rezultatelor pentru pacienți. Pentru mai multe informații despre acest caz, vizitați [acest link](#).

Principiile etice ale IA



Sursa imaginii | Freepik

Non-maleficență

Non-maleficența în IA se concentrează pe evitarea daunelor.

Non-maleficența se concentrează pe prevenirea daunelor, fie că acestea provin din acțiuni umane intenționate sau din comportamente neintenționate ale mașinilor, inclusiv influența neintenționată asupra comportamentului uman. Este vorba despre evitarea rezultatelor negative, indiferent de sursă.

(Floridi et al 2018)

Principiile etice ale IA



Sursa imaginii | Freepik

Exemplu privind principiul non-maleficenței

Poliția predictivă

- Imaginați-vă un algoritm AI utilizat în poliția predictivă care vizează în mod disproporționat comunitățile minoritare din cauza prejudecăților din datele pe care a fost antrenat. Acest sistem AI ar putea perpetua fără să vrea răul și nedreptatea, supunând persoane nevinovate unei supravegheri sau unui control sporit pe baza unor ipoteze greșite. În ciuda intenției sale de a reduce ratele criminalității, acțiunile AI pot duce la hărțuirea sau țintirea nejustificată a anumitor grupuri demografice.
- Utilizarea software-ului de poliție predictivă "Palantir" de către Departamentul de Poliție din New Orleans a vizat în mod disproporționat comunitățile minoritare, ceea ce a generat îngrijorări cu privire la perpetuarea prejudecăților și nedreptăților sistemice în practicile de aplicare a legii. Pentru mai multe informații despre acest caz, vizitați [acest link](#).

Principiile etice ale IA



Sursa imaginii | Freepik

Autonomie

Autonomia în IA implică un echilibru atent între procesul decizional uman și cel al mașinii, subliniind necesitatea de a păstra autodeterminarea umană.

- Autonomia presupune echilibrarea puterii decizionale între oameni și inteligența artificială. Aceasta susține valoarea intrinsecă a alegerii umane pentru deciziile importante, pledând pentru un model "decide-to-delegate" în care oamenii își păstrează autoritatea finală de a delega și, dacă este necesar, de a anula deciziile AI.

(Floridi et al 2018)

Principiile etice ale IA



Sursa imaginii | Freepik

Exemplu privind principiul autonomiei

Manipularea social media

- Un exemplu de AI care subminează autonomia ar putea fi un algoritm de recomandare pentru rețelele sociale care utilizează conținut personalizat pentru a manipula comportamentul și opiniile utilizatorilor fără consimțământul lor explicit. Prin alimentarea continuă a utilizatorilor cu conținut care se aliniază convingerilor sau prejudecăților lor existente, inteligența artificială restrânge expunerea utilizatorilor la diverse perspective și influențe, limitând astfel autonomia lor de a lua decizii în cunoștință de cauză pe baza unei game largi de informații.
- **Exemplu din viața reală:** Algoritmul News Feed al Facebook, conceput pentru a personaliza conținutul în funcție de interacțiunile utilizatorilor, a fost criticat pentru că a favorizat formarea de bule de filtrare și camere de ecou, amplificând potențial conținutul senzational sau polarizant și influențând comportamentul utilizatorilor fără consimțământul explicit. Pentru mai multe informații pe această temă, vizitați [acest link](#).

Principiile etice ale IA



Sursa imaginii | Freepik

Justiție

Principiul justiției în IA se referă la echitatea în procesul decizional și la distribuția socială a acestuia, cu scopul de a aborda disparitățile și de a promova echitatea.

- Principiul justiției se referă la utilizarea inteligenței artificiale pentru a aborda nedreptățile istorice, asigurând distribuirea echitabilă a beneficiilor inteligenței artificiale și prevenind noi prejudicii sau perturbări ale solidarității sociale.

(Floridi et al 2018)

Principiile etice ale IA



Sursa imaginii | Freepik

Exemplu privind principiul Justiției

Algoritmi de stabilire a sentințelor penale

- Algoritmii bazați pe inteligența artificială utilizați în stabilirea sentințelor penale pot perpetua în mod neintenționat prejudecățile împotriva anumitor grupuri demografice. În cazul în care datele istorice utilizate pentru antrenarea acestor algoritmi reflectă decizii părtinitoare luate de oameni, sistemul AI poate recomanda sentințe mai lungi sau pedepse mai aspre pentru persoanele din comunitățile marginalizate, exacerband inegalitățile existente în sistemul de justiție penală.
- S-a constatat că sistemul COMPAS, utilizat pentru evaluarea riscului în sentințele penale, prezintă prejudecăți rasiale, etichetând în mod disproporționat inculpații de culoare ca prezentând un risc ridicat și contribuind la pronunțarea unor sentințe nedrepte. Pentru mai multe informații despre acest caz, vizitați [acest link](#).

Megan T. Stevenson și Christopher Slobogin, University of Pennsylvania Law Review

Principiile etice ale IA



Sursa imaginii | Freepik

Explicabilitate

Explicabilitatea în inteligența artificială înseamnă să ne asigurăm că înțelegem modul în care inteligența artificială își ia deciziile și să fim capabili să explicăm clar aceste decizii. Aceasta garantează că sistemele AI sunt transparente și responsabile pentru acțiunile lor.

(Floridi et al 2018)

Principiile etice ale IA



Sursa imaginii | Freepik

Exemplu privind principiul Explicabilitate

Algoritmi de scoring al creditelor

- Să luăm în considerare un sistem de scoring al creditelor bazat pe inteligență artificială care determină bonitatea persoanelor, dar al cărui proces decizional este lipsit de transparentă. În cazul în care solicitanților li se refuză împrumuturi sau li se oferă condiții nefavorabile fără să înțeleagă factorii luați în considerare de algoritmul AI, capacitatea lor de a contesta sau de a face apel împotriva deciziilor este subminată. Fără explicații clare cu privire la modul în care inteligența artificială evaluează riscul de credit, persoanele fizice se pot simți nedreptățite și își pot pierde încrederea în integritatea sistemului.
- Breșa de date Equifax din 2017 a expus informațiile sensibile a milioane de consumatori, stârnind îngrijorări cu privire la lipsa de transparentă și securitate a algoritmilor de scoring al creditelor, evidențiind probleme mai largi ale industriei cu privire la vizibilitatea limitată a consumatorilor în ceea ce privește factorii care influențează bonitatea. Pentru mai multe informații Keith Frankish & William M. Ramsey despre acest caz, vizitați [acest link](#).

CADRE ETICE, ORIENTĂRI ȘI SETURI DE INSTRUMENTE ~~ALE IA~~



Cadre etice, orientări și seturi de instrumente ale

IA

Cadre etice

- **Cadrele etice merg dincolo de declarațiile de principii la nivel înalt. Ele oferă o abordare structurată pentru luarea deciziilor etice.**
- Multe cadre pentru inteligența artificială etică urmăresc să identifice potențialele provocări etice și să propună unele soluții pentru a depăși aceste provocări sau pentru a atenua riscurile asociate.
- Un cadru etic pentru IA servește drept structură fundamentală pentru elaborarea de legi, norme, standarde tehnice și bune practici în diferite sectoare și regiuni.
- Componentele tipice ale cadrelor etice sunt:
 - Concepte - relevante pentru dezbaterile aspectelor etice
 - Principiu - Principii etice (de exemplu, valori)
 - Preocupare - Modul în care principiile sunt amenințate prin utilizarea și dezvoltarea sistemelor AI
 - Remediu - Strategii, reguli și orientări pentru abordarea problemei

(Floridi & Cowls, 2019), (Prem, 2023: 701) (Ayling & Chapman, 2022)

Cadre etice, orientări și seturi de instrumente ale IA

Seturi de instrumente și orientări etice

- **Orientările etice oferă îndrumări practice și reguli specifice pentru un comportament etic în diverse contexte.**
- **Seturile de instrumente și orientările etice** modelează inovarea bazată pe IA și sprijină **aplicarea practică a principiilor etice ale IA.**
- Seturile de instrumente și orientările explică modul de aplicare a principiilor etice în proiectarea, implementarea și utilizarea inteligenței artificiale.
- Cadrele, orientările și seturile de instrumente se suprapun frecvent în utilizarea lor.
- În paginile următoare, sunt prezentate mai detaliat două exemple de cadre: cadrul HLEG pentru inteligența artificială demnă de încredere și cadrul OCDE de clasificare a inteligenței artificiale

(Zhou et al., 2020)

Cadre etice, orientări și seturi de instrumente ale

IA

Grupul de experți la nivel înalt privind inteligența artificială; inteligență artificială demnă de încredere



- În ciuda acordului că IA ar trebui să fie etică, există dezbateri cu privire la ceea ce constituie "IA etică" și ce cerințe etice și standarde tehnice sunt necesare pentru realizarea acesteia.
- Comisia Europeană a prezentat o strategie pentru dezvoltarea IA în Europa, care pune accentul pe IA etică, sigură și avansată.
- Orientările etice pentru o IA demnă de încredere au fost prezentate de AI-HLEG în aprilie 2019 pentru a promova o IA demnă de încredere.
- Inteligența artificială **demnă de încredere** este recunoscută ca fiind obiectivul esențial pentru promovarea încrederii în dezvoltarea și aplicarea inteligenței artificiale, bazată pe un cadru solid care să asigure fiabilitatea acesteia.
- Vezi mai multe de la <https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai>

(Leijnen et al., 2020), HLEG (2019). Grupul de experți la nivel înalt privind inteligența artificială. Orientări etice pentru o inteligență artificială demnă de încredere. Bruxelles: Comisia Europeană

Cadre etice, orientări și seturi de instrumente ale IA

IA de încredere - cadru

Inteligența artificială demnă de încredere are trei componente, care ar trebui îndeplinite pe parcursul întregului ciclu de viață al sistemului:

- Legal - respectând toate legile și reglementările aplicabile
- Etic - respectarea principiilor și valorilor etice
- Robust - atât din punct de vedere tehnic, cât și ținând cont de mediul său social

Cadre etice, orientări și seturi de instrumente ale

IA Grupul de experți la nivel înalt privind inteligența artificială: cadru pentru inteligența artificială demnă de încredere

Bazele inteligenței artificiale demne de încredere

- Patru **principii etice** bazate pe drepturile fundamentale

Realizarea inteligenței artificiale demne de încredere

- Punerea în aplicare a șapte **cerințe-cheie**

Evaluarea inteligenței artificiale demne de încredere

- Pentru operaționalizarea cerințelor-cheie, adaptați **lista de evaluare** la aplicația AI specifică

HLEG (2019). Grupul de experți la nivel înalt privind inteligența artificială. Orientări etice pentru o inteligență artificială demnă de încredere. Bruxelles: Comisia Europeană.

Cadre etice, orientări și seturi de instrumente ale

IA

Fundamentele inteligenței artificiale demne de încredere: principii etice

Respectarea autonomiei umane

- IA ar trebui să sporească autodeterminarea, să evite manipularea și să sprijine abilitățile cognitive, sociale și culturale ale omului. Ea trebuie să permită supravegherea umană și alegerea semnificativă și să sprijine oamenii la locul

Prevenirea daunelor

- Inteligența artificială nu trebuie să provoace daune, trebuie să protejeze demnitatea umană, să fie sigură și robustă și să acorde o atenție specială grupurilor vulnerabile și mediului. Aceasta ar trebui să prevină efectele

Echitate

- AI ar trebui să asigure distribuția echitabilă a beneficiilor și costurilor, să prevină prejudecățile și discriminarea și să promoveze egalitatea de șanse. Aceasta necesită un echilibru între mijloace și scopuri și permite contestarea

Explicabilitate

- Procesele IA trebuie să fie transparente, cu o comunicare clară a capacităților și scopurilor. Atunci când explicarea completă nu este posibilă, alte măsuri precum trasabilitatea și auditabilitatea ar trebui să asigure respectarea drepturilor.

- AI HLEG enumeră patru principii etice, înrădăcinate în drepturile fundamentale, care trebuie respectate pentru a garanta că sistemele AI sunt dezvoltate, implementate și utilizate într-un mod demn de încredere.
- Principiile etice sunt specificate ca imperative etice, astfel încât practicienii AI ar trebui să depună întotdeauna eforturi pentru a le respecta.
- De asemenea, AI HLEG recunoaște și abordează tensiunile potențiale dintre aceste principii.

<https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Cadre etice, orientări și seturi de instrumente ale

IA Realizarea unei IA demne de încredere: cerințe-cheie

- Principiile etice stabilite trebuie transformate în **cerințe** definitive pentru realizarea unei IA demne de încredere. Aceste cerințe se referă la toți participanții la ciclul de viață al sistemului IA:
- Dezvoltatori: persoane sau echipe care desfășoară activități de cercetare, proiectare sau dezvoltare în domeniul IA.
- Deployeri: organizații care implementează inteligența artificială în operațiunile lor și oferă produse sau servicii bazate pe inteligența artificială.
- Utilizatori finali: persoanele care interacționează cu sistemele AI, direct sau indirect.

Agenție umană și supraveghere

- drepturile fundamentale, agenția umană și supravegherea umană

Robustețe tehnică și siguranță

- rezistență la atac și securitate, plan de rezervă și siguranță generală, acuratețe, fiabilitate și reproductibilitate

Confidențialitatea și guvernanta datelor

- respectarea vieții private, a calității și integrității datelor, precum și a securității datelor

Transparență

- trasabilitate, explicabilitate și comunicare

Diversitate, nediscriminare și echitate

- evitarea prejudecăților nedrepte, accesibilitatea și proiectarea universală, precum și participarea părților interesate

Bunăstarea socială și de mediu

- durabilitate și respect față de mediu, impact social, societate și democrație

Responsabilitate

- auditabilitatea, minimizarea și raportarea impactului negativ, compromisurile și despăgubirile.

HLEG (2019). Grupul de experți la nivel înalt privind inteligența artificială. Orientări etice pentru o inteligență artificială demnă de încredere. Bruxelles: Comisia Europeană

Cadre etice, orientări și seturi de instrumente ale

IA Evaluarea inteligenței artificiale demne de încredere: lista de evaluare ALTAI

- Pentru a contribui la dezvoltarea unei IA demne de încredere, a fost dezvoltat un instrument care transformă aceste principii ale IA în măsuri practice.
- Cunoscut sub numele de Lista de evaluare pentru AI de încredere (ALTAI), acest instrument oferă o listă de verificare cuprinzătoare și dinamică care servește drept foaie de parcurs pentru dezvoltatori și implementatori în vederea integrării acestor principii în aplicațiile lor AI.
- ALTAI facilitează acest proces prin oferirea unui set de pași concreți pentru autoevaluare, asigurându-se astfel că tehnologiile IA sunt dezvoltate într-un mod care maximizează beneficiile utilizatorilor, reducând în același timp expunerea la riscuri inutile.
- Acces la instrumentul ALTAI



SURSĂ IMAGINE | ALTAI

(Grupul de experți la nivel înalt privind inteligența artificială (AI HLEG), 2020)

Cadre etice, orientări și seturi de instrumente ale IA

Cadrul de clasificare a inteligenței artificiale al OCDE

- Cadrul de clasificare a IA al OCDE, creat de rețeaua de experți OECD.AI, este conceput pentru a asista diverse părți interesate în evaluarea potențialului și riscurilor diverselor sisteme de IA.
- Acest instrument facilitează elaborarea strategiilor în materie de IA și vizează menținerea coerenței politicilor la nivel internațional.
- Cadrul se bazează pe principiile IA ale OCDE, susținând valori precum corectitudinea, transparența, siguranța și responsabilitatea, precum și pledând pentru dezvoltarea capacității umane și colaborarea internațională.



Sursa imaginii | Cadrul OCDE

<https://www.oecd.org/digital/artificial-intelligence/>

UE AI ACT



Legea UE privind inteligența artificială

- Legea privind inteligența artificială este un proiect de lege al UE privind inteligența artificială | Legea privind inteligența artificială - primul de acest fel din lume. Se preconizează că legea va fi pusă în aplicare în 2026.
- Aceasta se aplică dezvoltării, implementării și utilizării IA în UE sau atunci când aceasta va afecta persoanele din UE.
- Comisia Europeană a publicat propunerea de regulament privind inteligența artificială (AI Act) la 21 aprilie 2021. Actul este primul cadru juridic care stabilește norme armonizate pentru dezvoltarea, introducerea pe piață și utilizarea inteligenței artificiale în Uniunea Europeană.



SURSĂ IMAGINE | [The AI Act Explorer](#) | [Legea UE privind inteligența](#)

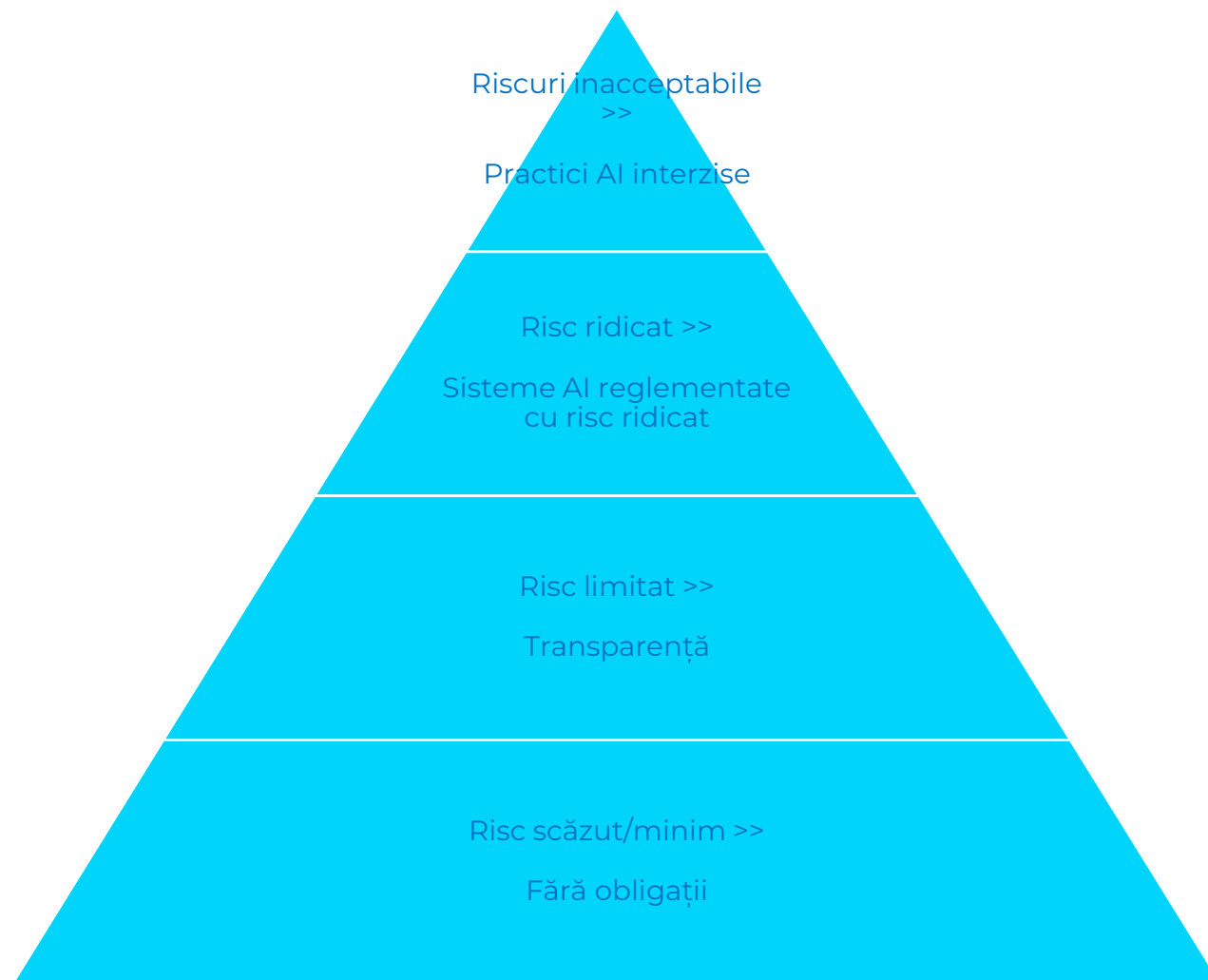
[artificială
https://artificialintelligenceact.com](https://artificialintelligenceact.com)
<https://www.euaiact.com>

Legea UE privind inteligența artificială

Cadru bazat pe riscuri

Principiile etice stabilite trebuie transformate în **cerințe** definitive pentru realizarea unei IA demne de încredere. Aceste cerințe se referă la toți participanții la ciclul de viață al sistemului IA:

- **Dezvoltatori:** persoane sau echipe care desfășoară activități de cercetare, proiectare sau dezvoltare în domeniul IA.
- **Deployeri:** organizații care implementează inteligența artificială în operațiunile lor și oferă produse sau servicii bazate pe inteligența artificială.
- **Utilizatori finali:** persoanele care interacționează cu sistemele AI, direct sau indirect.
- **Societatea în sens larg:** cei care sunt afectați de IA, atât direct, cât și indirect.



"Piramida riscurilor" din Legea UE privind IA (Sursa: Parlamentul European)

Legea UE privind inteligența artificială

Cadru bazat pe riscuri | Risc minim sau inexistent



aplicațiile AI care sunt considerate a prezenta **un risc minim sau inexistent** pentru drepturile sau siguranța cetățenilor. Marea majoritate a sistemelor de inteligență artificială se încadrează în categoria de risc minim.

Exemple de aplicații ale IA cu risc minim:

- Sisteme de recomandare bazate pe inteligență artificială
- Filtre de spam

În conformitate cu reglementările actuale, aceste sisteme nu sunt supuse unor obligații specifice noi, ci trebuie să respecte legile preexistente.

Legea UE privind inteligența artificială

Cadru bazat pe riscuri | Risc limitat



Sistemele AI cu **risc limitat** necesită transparență.

Pentru aceste sisteme AI, cerințele se referă la transparență: Utilizatorii ar trebui să fie informați că interacționează cu inteligența artificială și să dispună de informațiile necesare pentru a decide cu privire la continuarea utilizării.

Exemple de aplicații AI cu risc limitat:

- Inteligență artificială care generează sau manipulează imagini sau conținut audio.
- AI care creează sau modifică conținutul video, cum ar fi deepfakes.

Legea UE privind inteligența artificială

Cadru bazat pe riscuri | Risc ridicat



Sistemele AI cu risc ridicat pot avea un impact semnificativ asupra șanselor de viață ale unui utilizator.

Aceste sisteme trebuie să respecte cerințe stricte înainte de a fi implementate pe piața UE, inclusiv obligații privind gestionarea riscurilor și guvernanta datelor.

Exemplele de IA cu risc ridicat includ:

- Gestionarea și exploatarea infrastructurii critice
- Educație și formare profesională
- Ocuparea forței de muncă, gestionarea lucrătorilor și accesul la activități independente
- Accesul la serviciile private esențiale și la serviciile și prestațiile publice
- Aplicarea legii
- Migrația, azilul și gestionarea controlului la frontieră
- Asistență în interpretarea juridică și aplicarea legii.

Legea UE privind inteligența artificială

Cadru bazat pe riscuri | Risc inacceptabil



Sistemele AI cu **risc inacceptabil** sunt considerate a avea o amenințare clară la adresa drepturilor fundamentale ale oamenilor sunt interzise. Sistemele cu risc inacceptabil sunt cele care exploatează, manipulează sau utilizează tehnici subliminale.

Vânzarea acestor sisteme AI este interzisă pe piața UE.

Exemple de utilizare interzisă a AI includ:

- Manipularea cognitiv-comportamentală a persoanelor sau a grupurilor vulnerabile specifice: de exemplu, jucării activate vocal care încurajează comportamentele periculoase ale copiilor
- Punctaj social: clasificarea persoanelor în funcție de comportament, statut socio-economic sau caracteristici personale
- Identificarea și clasificarea biometrică a persoanelor
- Sisteme de identificare biometrică în timp real și la distanță, cum ar fi recunoașterea facială

CICLUL DE VIAȚĂ AL DEZVOLTĂRII AI



Ciclul de viață al dezvoltării AI

Transformarea principiilor, cadrelor și orientărilor etice în acțiuni concrete

Importanța principiilor, cadrelor și liniilor directoare în dezvoltarea IA este bine recunoscută, însă aplicarea lor practică în timpul creării soluțiilor IA prezintă mai multe **obstacole**:

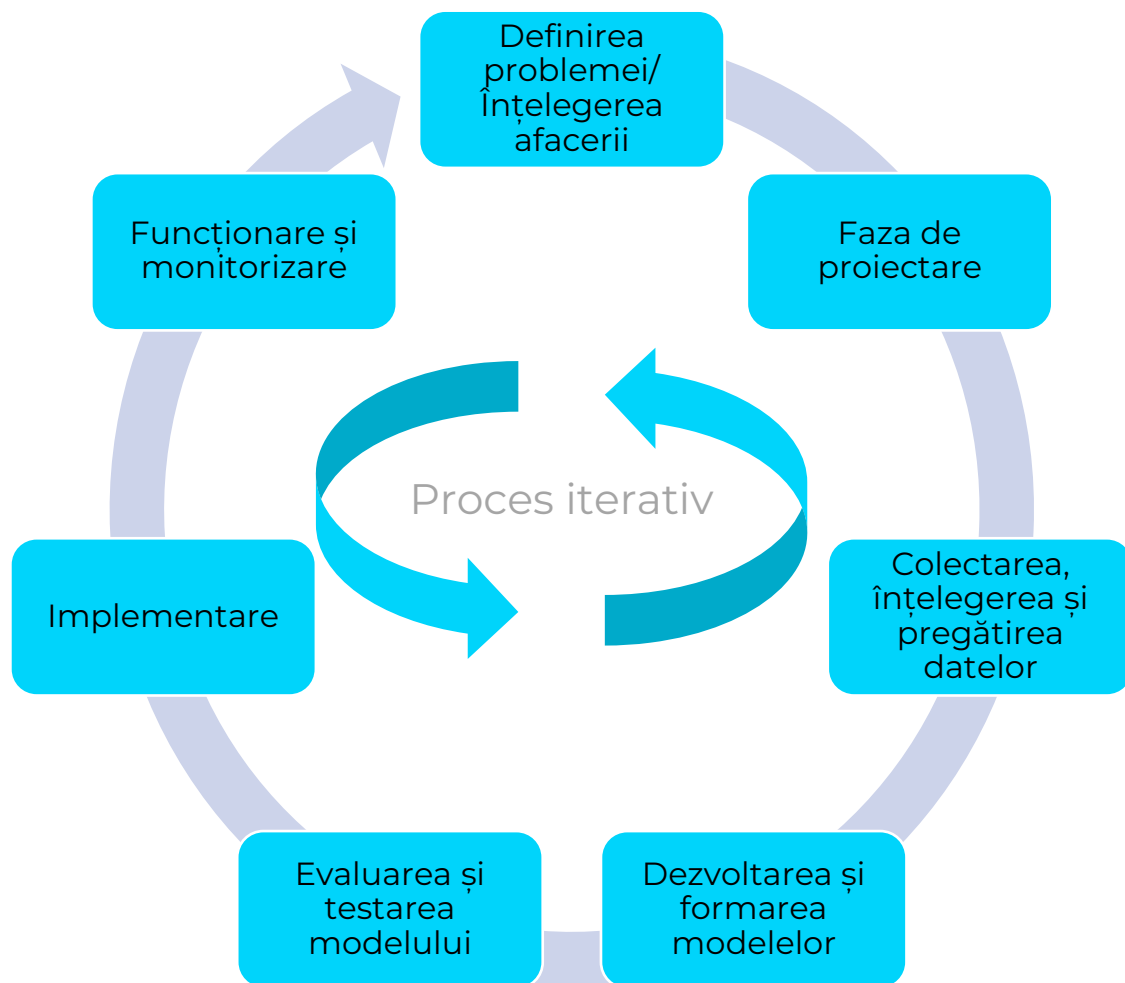
- Principiile IA tind să fie abstracte, diverse și complexe.
- Conversia comportamentului uman nuanțat în instrumente simple și universale este complexă.
- Principiile sunt adesea supuse interpretării personale, influențate de trecutul fiecăruia.
- Există o absență a standardizării universale a reglementărilor, cadrelor și orientărilor (chiar dacă au fost făcute unele generalizări, așa cum am prezentat anterior)
- Deși cadrele sunt utile pentru identificarea problemelor, este posibil ca acestea să nu ofere îndrumări concrete cu privire la pașii specifici de urmat

Ciclul de viață al dezvoltării AI

Transformarea principiilor, cadrelor și orientărilor etice în acțiuni concrete: Ciclul de viață al dezvoltării IA

- Instrumentele disponibile concepute pentru a aborda provocările etice în soluțiile AI sunt accesibile, dar adesea acoperă doar un număr restrâns de principii sau etape de dezvoltare a soluțiilor AI
- Pentru a transforma principiile etice în acțiuni concrete, se recomandă să se privească procesul de dezvoltare a IA ca pe o succesiune de etape care începe cu un caz de afaceri și se încheie cu modul operațional al soluției. (Ciclul de viață al dezvoltării IA)
- Ciclul de viață al dezvoltării IA, adoptă o viziune holistică asupra considerentelor etice în toate etapele de dezvoltare

Ciclul de viață al dezvoltării AI



- Ca orice proiect, proiectul de design AI este recomandat să urmeze pașii de proiectare pentru a transforma ideile inovatoare în soluții practice, bine definite
- Este important să înțelegeți că fiecare etapă joacă un rol esențial în dezvoltarea soluției, formând un ciclu de viață al dezvoltării AI
- Procesul este ITERATIV, ceea ce înseamnă că ar putea fi necesar să se revină la etapa anterioară în caz de nevoie și să se modifice soluția

Etapale de proiectare a dezvoltării IA prezentate aici sunt o combinație a diferitelor modele descrise de articolele Data Science Process Alliance (2023), Morley et.al' (2019), Prehm's (2022) și Rochel et.al'a (2020).

Alianța pentru procesul științei datelor /Morley (2019) / Prem (2023) / Rochel & Evequoz (2020)



- Etapele de proiectare a dezvoltării IA prezentate aici sunt o combinație a diferitelor modele descrise de articolele Data Science Process Alliance (2023), Morley et.al' (2019), Prehm's (2022) și Rochel et.al'a (2020).

CU1 | AI ethics - a practical approach

Ciclul de viață al dezvoltării AI



DEFINIREA
PROBLEMEI ȘI
ÎNȚELEGEREA
AFACERII

Definirea problemei este esențială pentru înțelegerea nevoilor clienților și a rolului soluției AI.



FAZA DE
PROIECTARE

Faza de proiectare transformă cerințele de afaceri în cerințe de proiectare pentru ingineri



ÎNȚELEGEREA ȘI
PREGĂTIREA
COLECTĂRII
DATELOR

Colectarea, înțelegerea și pregătirea datelor constituie fundamentul soluției



DEZVOLTAREA ȘI
FORMAREA
MODELELOR

Dezvoltarea și evaluarea modelului au ca scop transformarea datelor într-o soluție fezabilă care să abordeze problema inițială



EVALUAREA ȘI
TESTAREA
MODELULUI

Modelul trebuie testat pe scară largă, iar funcționalitatea este reflectată în nevoile clientului



DEPLOIERE

Implementarea constă în aducerea soluției la utilizatori



FUNCȚIONARE
ȘI
MONITORIZARE

Soluția trebuie monitorizată în timp și întreținută în funcție de evoluția nevoilor

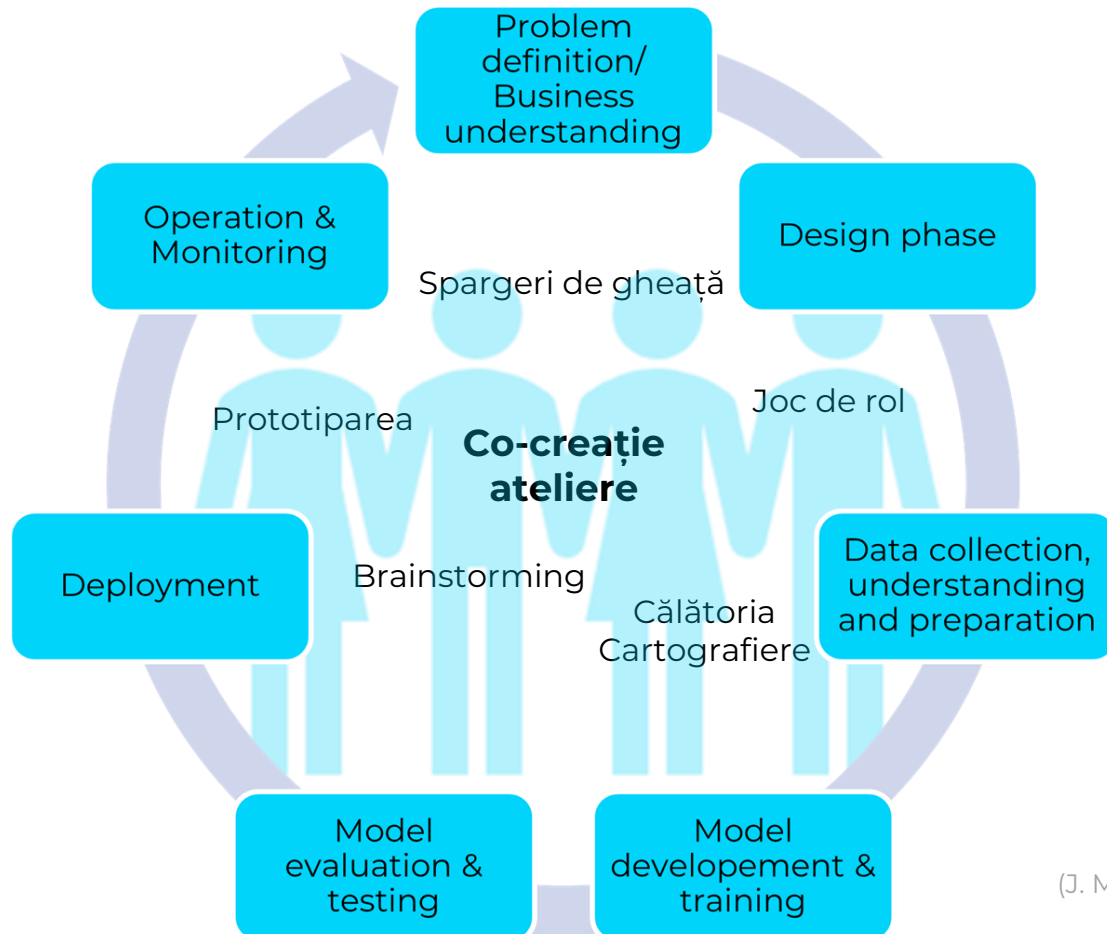
IMPLICAREA PĂRȚILOR INTERESATE ÎN DEZVOLTAREA SOLUȚIILOR AI



Sursa imaginii | Freepik

Implicarea părților interesate în dezvoltarea soluțiilor AI

De ce este important să vă angajați diverse părți interesate în dezvoltarea soluțiilor AI?

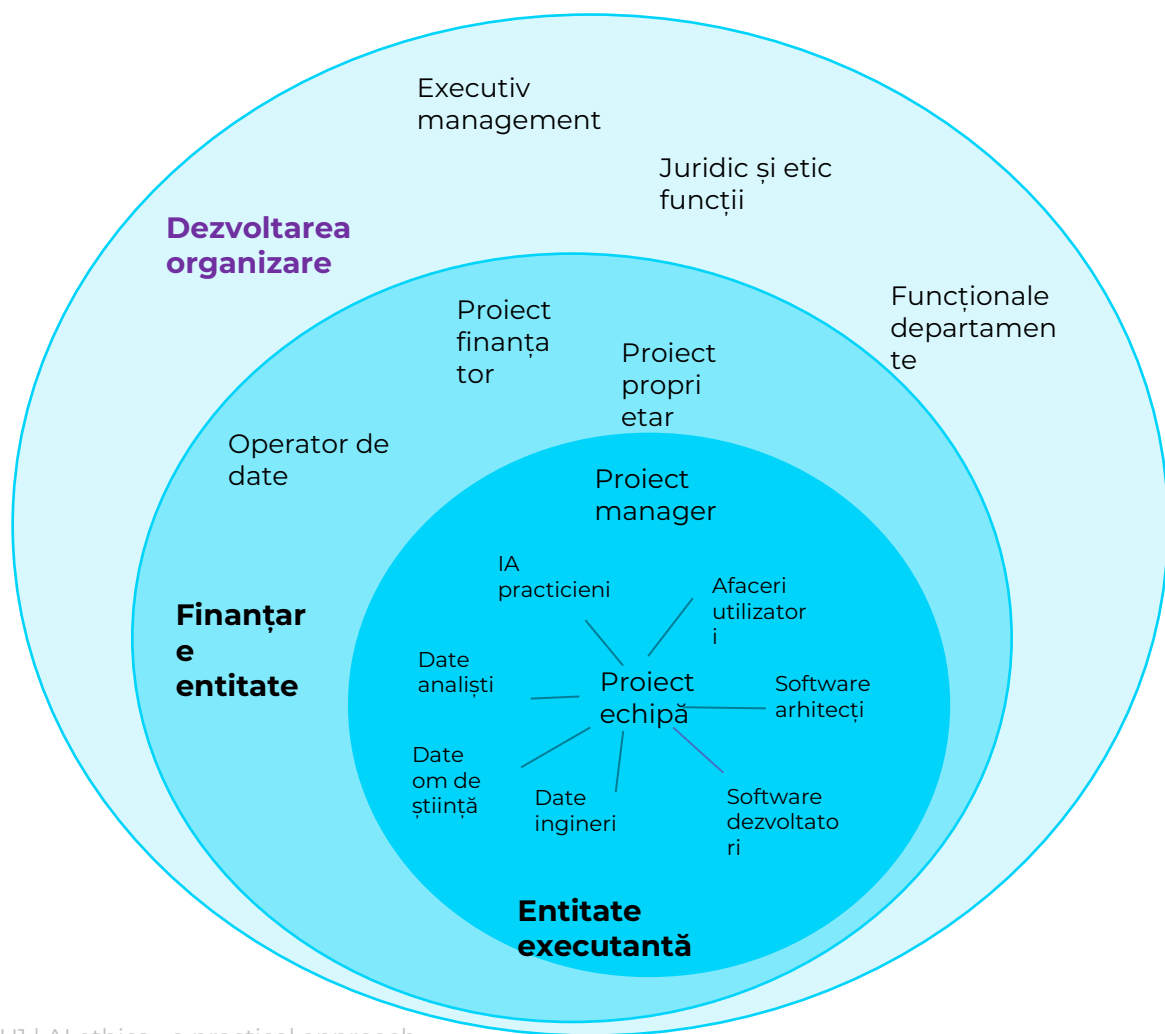


- Pentru a se asigura că impactul dezvoltării AI este luat în considerare din diferite puncte de vedere și pentru a dezvolta sisteme demne de încredere, este esențial să se implice și să se consulte mai multe părți interesate în procesul de dezvoltare, inclusiv pe cei care vor fi afectați de soluția AI.
- Co-crearea este o practică utilizată pentru colaborarea cu părțile interesate, inclusiv utilizatorii finali sau clienții, în timpul unui proces de proiectare, pentru a împărtăși informații între părțile interesate cu roluri și expertiză diferite și permite o mai bună înțelegere a nevoilor utilizatorilor finali și, astfel, proiectarea de soluții care iau în considerare diferiți utilizatori.

(J. Morley et al., 2019) / (ONESCO, 2023) / (Interaction Design Foundation, Co-creation) / (Anika Sahin, 2020)

Implicarea părților interesate în dezvoltarea soluțiilor AI

Părțile interesate de dezvoltare

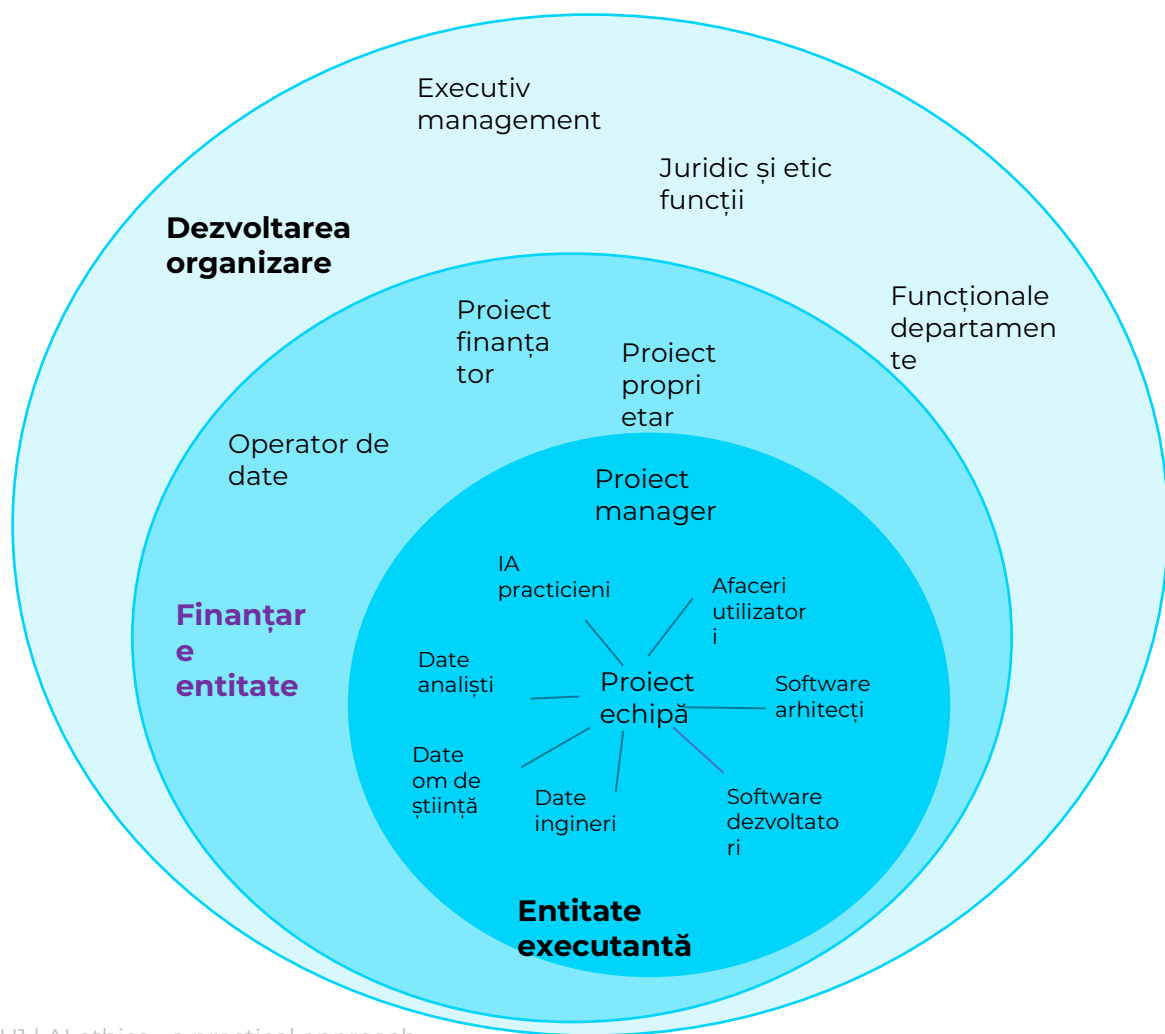


Dezvoltarea organizației

- Sponsorizează proiectul IA și stabilește domeniul de aplicare al proiectului.
- Rolurile proiectului sunt **managementul executiv, departamentele funcționale și funcțiile juridice și etice**:
 - **Conducerea executivă** este responsabilă pentru deciziile strategice, financiare și de dezvoltare ale companiei.
 - **Departamentul funcțional** este o unitate organizațională specializată din cadrul unei companii care se concentrează pe un set specific de sarcini sau activități, adesea grupate pe baza unor funcții comune, cum ar fi finanțele, marketingul sau resursele umane.
 - **Funcțiile juridice și etice** din cadrul unei organizații implică diverse roluri și responsabilități pentru a asigura conformitatea cu legile, reglementările și standardele etice. (Miller, 2022)

Implicarea părților interesate în dezvoltarea soluțiilor AI

Părțile interesate de dezvoltare



Entitatea finanțatoare

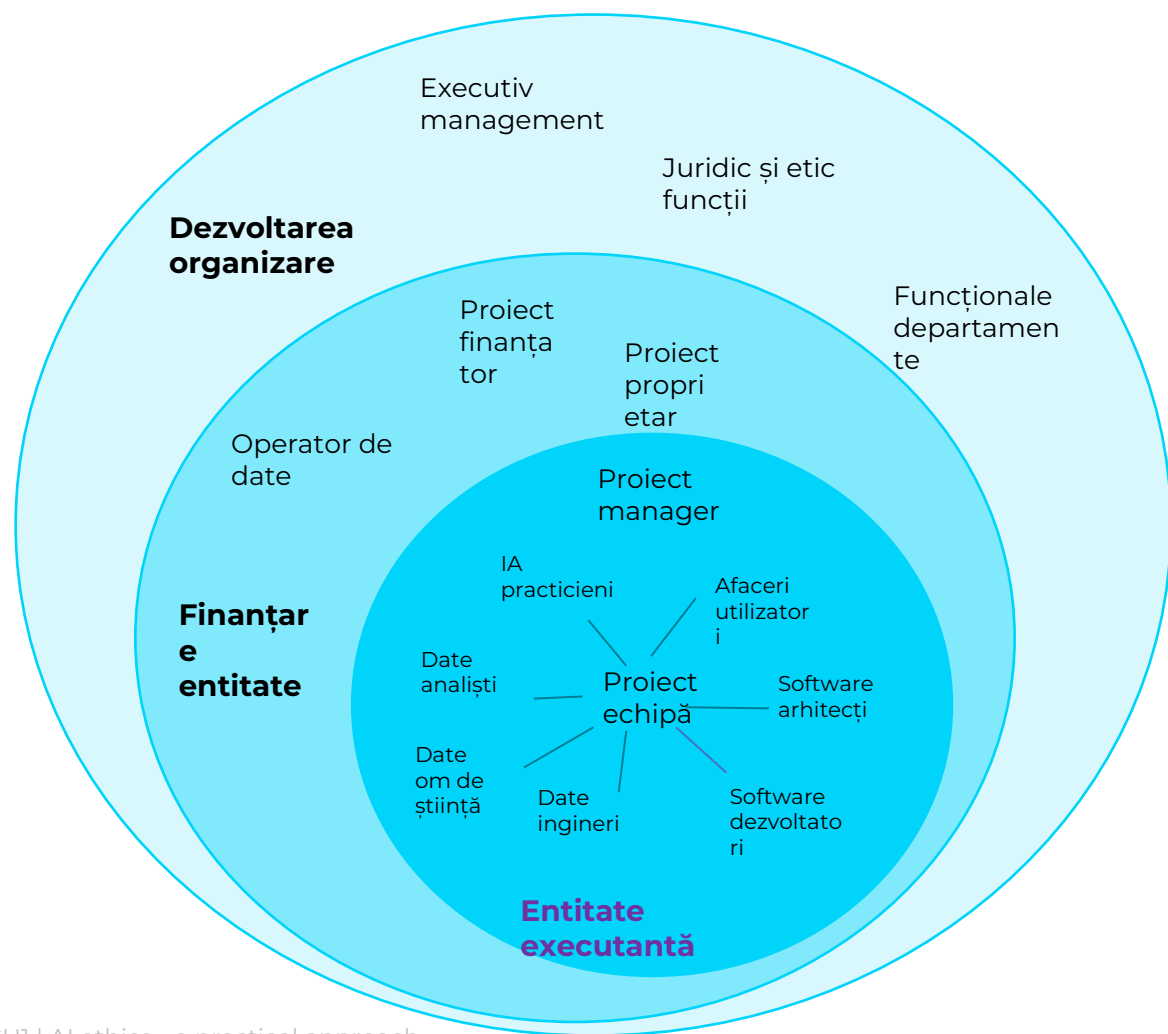
Rolurile proiectului sunt **finanțator**, **proprietar** și **operator de date**.

- **Finanțatorul proiectului finanțează** proiectul și alocă resursele pentru proiect.
- **Proprietarul proiectului** oferă direcție strategică.
- **Operatorul de date** este responsabil de gestionarea datelor și de utilizarea acestora în dezvoltarea sistemului.

(G.J. Miller, 2022)

Implicarea părților interesate în dezvoltarea soluțiilor AI

Părțile interesate de dezvoltare



Entitate executantă

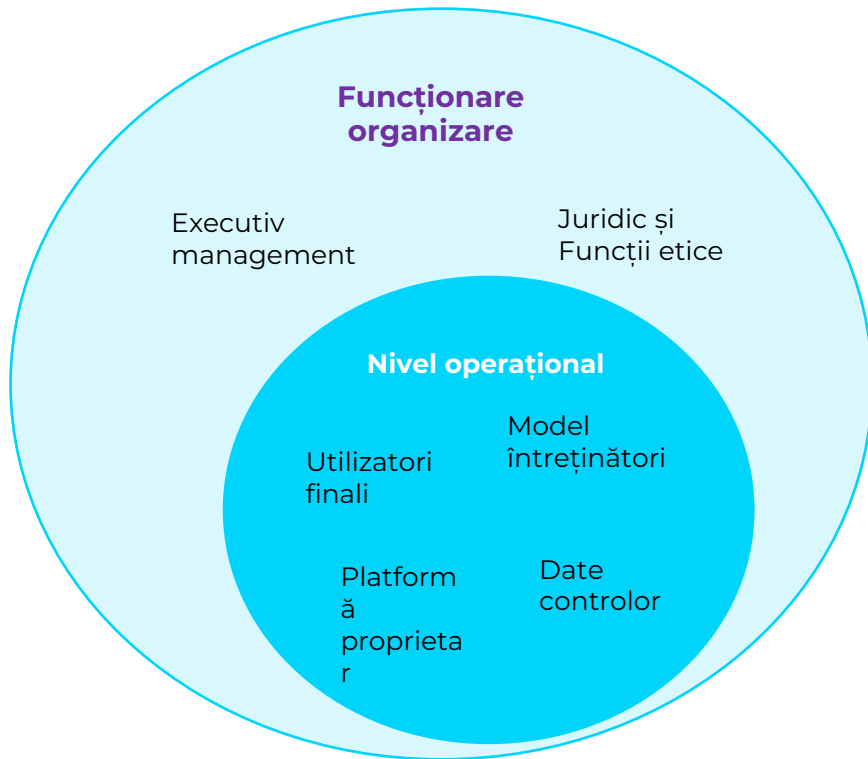
Rolurile proiectului sunt **managerii de proiect** și **echipa de proiect**.

- **Managerul de proiect** este responsabil pentru rezultatele proiectului și se asigură că standardele și așteptările în materie de etică, confidențialitate și securitate sunt respectate și că părțile interesate relevante sunt implicate.
- **Echipa de proiect AI este** formată din arhitecți software, dezvoltatori software, ingineri de date, cercetători de date, analiști de date, practicieni AI și utilizatori de afaceri.

(G.J. Miller, 2022)

Implicarea părților interesate în dezvoltarea soluțiilor AI

Părțile interesate de utilizare



Organizare operațională

- O organizație operațională este o parte interesată a clientului.
- La fel ca în organizația de dezvoltare, rolurile de proiect ale organizației operaționale includ **managementul executiv** și **funcțiile juridice și etice**.
- La nivel operațional, rolurile proiectului sunt **utilizatori finali**, **responsabili cu menținerea modelelor**, **controlori de date** și **proprietari de platforme**.
 - **Utilizatorii finali** sunt persoane sau grupuri care interacționează cu sistemul de inteligență artificială prin contracte de muncă sau de servicii cu organizația de exploatare sau în calitate de consumatori.
 - **Responsabilii cu menținerea modelelor** sunt responsabili pentru gestionarea, actualizarea și optimizarea continuă a modelelor de învățare automată utilizate în sistemul de inteligență artificială.
 - **Operatorii de date** stabilesc scopurile și mijloacele de prelucrare a datelor cu caracter personal în cadrul sistemului AI.
 - **Proprietarii platformei** sunt responsabili pentru infrastructura

Implicarea părților interesate în dezvoltarea soluțiilor AI

Părțile interesate externe



Persoane și societate

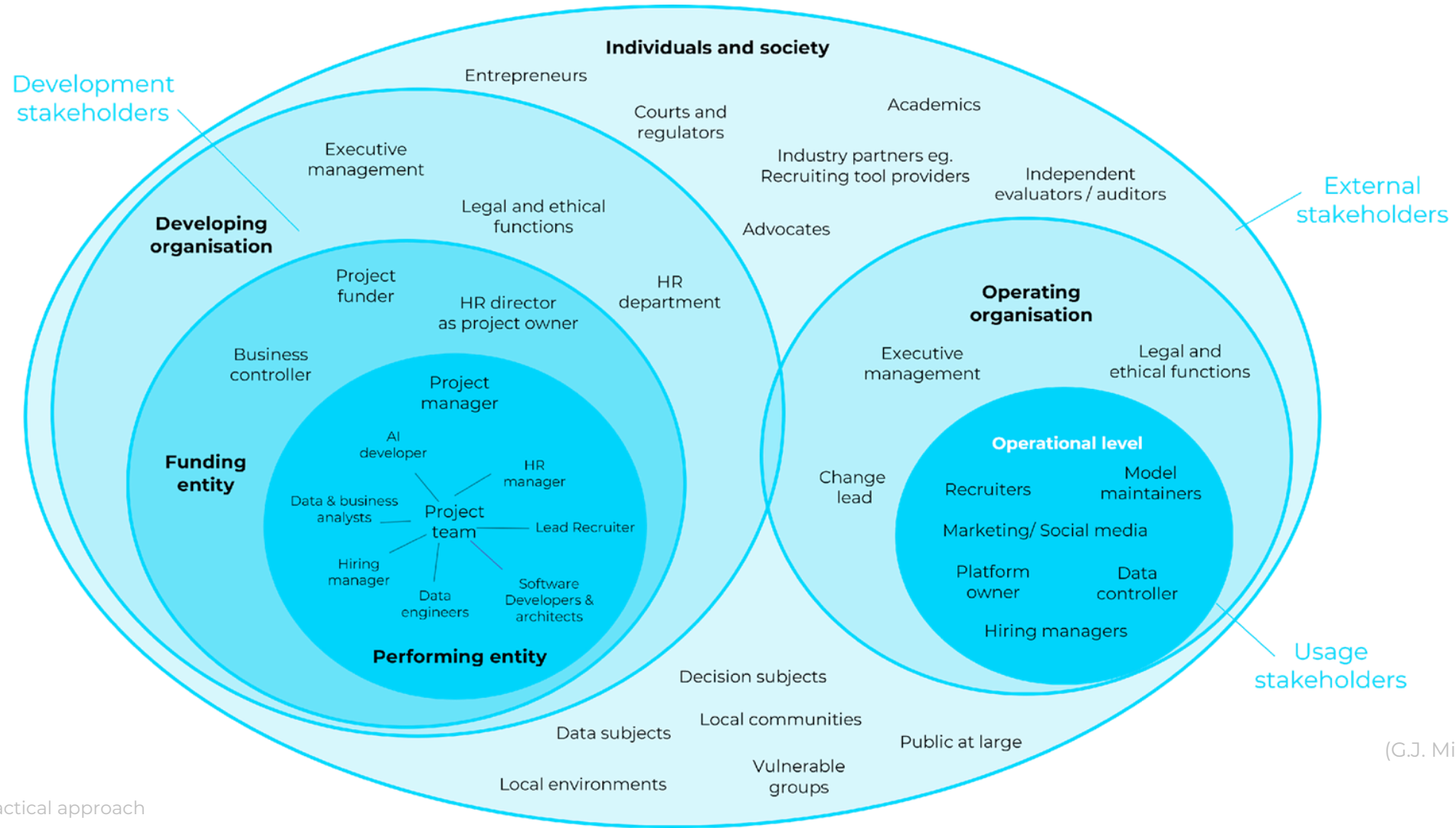
- **Subiecții individuali** sunt persoanele care au încheiat acorduri cu organizația de dezvoltare sau de exploatare, cum ar fi contractele de servicii sau de muncă. Acest grup include persoanele vizate, persoanele care iau decizii și lucrătorii și sunt considerate părți interesate pasive.
- **Societatea este** formată din indivizi și din public. Aceștia ar putea fi afectați negativ de dezvoltarea sau utilizarea sistemului de inteligență artificială.
- Acest grup cuprinde comunități locale, medii și grupuri vulnerabile precum persoanele cu handicap, grupurile minoritare, minorii și publicul larg.

Inginerii AI fac parte dintr-o discuție societală mai amplă, în care trebuie să joace un rol activ în integrarea altor perspective.

- **Reprezentanții** sunt grupuri și organizații care nu sunt afectate de sistemul IA, dar care ar putea reprezenta alte părți în cadrul proiectului.
- **Reprezentanții formali** includ instanțele, legile, autoritățile de reglementare și sindicatele.
- **Antreprenorii** pot fi implicați în dezvoltarea sistemului AI.
- **Evaluatorii independenți** sunt persoane sau grupuri din exterior, de exemplu reporteri, auditori sau revizori.

Implicarea părților interesate în dezvoltarea soluțiilor AI

Părțile interesate în dezvoltarea IA

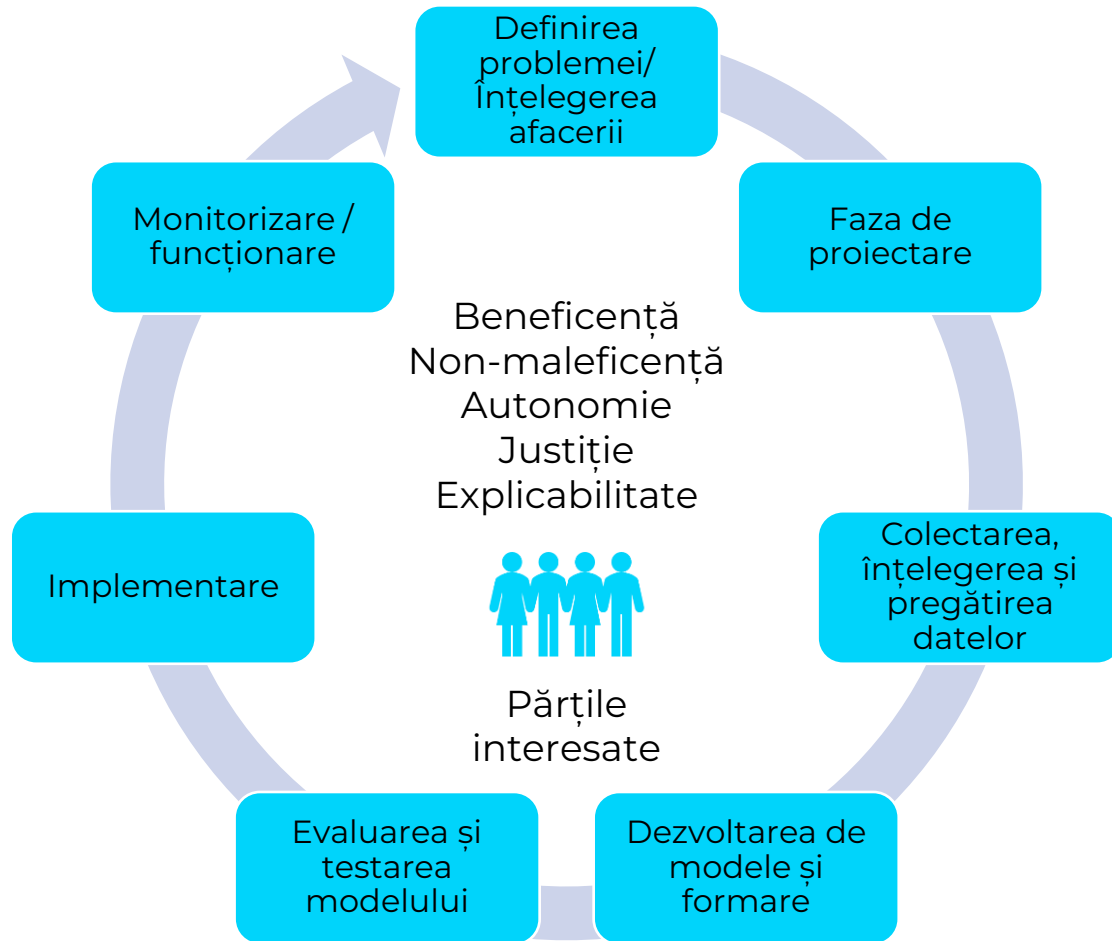


(G.J. Miller, 2022)

PRINCIPII ETICE ȘI IMPLICAREA PĂRȚILOR INTERESATE ÎN CICLUL DE VIAȚĂ AL DEZVOLTĂRII AI



Principii etice și părți interesate în ciclul de viață al dezvoltării IA



Pe tot parcursul ciclului de viață al IA, este important să ne concentrăm asupra aspectelor etice și asupra implicării părților interesate.

Următoarele pagini prezintă ciclul de viață din trei perspective:

- 1. Descrierea etapei**
- 2. Principalele principii etice**
- 3. Părțile interesate implicate**

Nu toate principiile etice au aceeași greutate în orice context sau în orice etapă a dezvoltării și, prin urmare, în cele ce urmează sunt prezentate doar principalele considerații și puncte de vedere.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Definirea problemei și înțelegerea afacerii



Prima etapă crucială pentru un proiect de succes

- Înțelegeți problema de rezolvat sau oportunitatea de investigat.
- Efectuați cercetări și interviuri cu părțile interesate pentru a formula o enunțare clară a problemei.
- Înțelegerea nevoilor clienților și a modului în care acestea se traduc într-o soluție de afaceri viabilă.
- Definiți obiectivele și criteriile de succes ale soluției.
- Efectuați un studiu de fezabilitate pentru a determina dacă IA este cea mai potrivită soluție.
- Evaluați fezabilitatea tehnică și etică a proiectului

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Definirea problemei și înțelegerea afacerii

Principii etice de luat în considerare



Stabilirea bazei - abordarea problemelor etice de bază, cum ar fi drepturile omului, impactul asupra mediului și preocupările societale.

Beneficență - asigurați-vă că soluția AI sporește bunăstarea individuală și binele societății. Domenii cheie de luat în considerare.

- Participarea **părților interesate**
 - Crearea unor sisteme de încredere care să sprijine înflorirea umană.
 - Implicarea persoanelor afectate în procesul de dezvoltare.
- **Protecția** drepturilor **fundamentale**
 - Protejarea drepturilor individuale, consolidând integritatea etică.
- Sustenabilitatea **mediului**
 - Minimizarea impactului asupra mediului prin practici durabile.
- **Justificare**
 - Definirea clară și corelarea scopului sistemului cu beneficii societale tangibile.

Morley (2019) / Prem (2023)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Definirea problemei și înțelegerea afacerii

Principii etice de luat în considerare



Non-maleficență

- Asigurați-vă că identificarea problemei și soluția propusă nu dăunează persoanelor sau comunităților.
- Luați în considerare cu atenție și monitorizați impactul sistemului asupra bunăstării fizice și mentale.
- Evaluarea potențialelor riscuri de siguranță și a strategiilor de reducere a acestora.

Justiție

- Evaluați impactul sistemului asupra instituțiilor, democrației și societății.
- Luați în considerare dacă anumite grupuri sunt favorizate sau dezavantajate.
- Examinați dacă proiectul ar putea încălca autonomia individuală.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Definirea problemei și înțelegerea afacerii Părțile interesate



- **Utilizatori finali:** utilizatorii principali ai soluției AI
- **AI/data scientists:** formulează problema și conceptualizează o soluție
- **Comunități** afectate
- **Experți în domeniu:** cunoaștere aprofundată a industriei în care se aplică IA
- **Dezvoltatorii AI:** implicarea timpurie le aduce o înțelegere profundă a problemei
- **Managerii de proiect:** supraveghează întreaga desfășurare a proiectului
- **Părțile interesate din mediul de afaceri:** asigurarea alinierii la obiectivele și resursele de afaceri
- **Eticienii și oamenii de știință din domeniul social:** evidențierea impactului potențial

(D. De Silva & D. Alahakoon, 2022)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Faza de proiectare



Această etapă transformă munca de bază efectuată în timpul etapei de definire a problemei în planuri detaliate pentru construirea și implementarea sistemului AI

- Specificați ce urmărește proiectul să realizeze (obiective și rezultate)
- Evaluați costurile financiare în raport cu beneficiile anticipate.
- Prezentați funcționalitățile pe care trebuie să le aibă sistemul de inteligență artificială.
- Decideți cu privire la abordarea pentru utilizarea datelor, cum ar fi utilizarea modelelor pre-antrenate.
- Stabiliți parametri pentru a măsura succesul proiectului.
- Integrarea orientărilor etice pentru a reglementa procesul de dezvoltare.
- Elaborarea unui calendar și gestionarea implicării părților interesate.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Faza de proiectare

Principii etice de luat în considerare



Justiție | Evitarea prejudecăților nedrepte

- **Cerințe de proiectare**

- Implementarea mecanismelor de identificare și atenuare a prejudecăților din date și algoritmi.
- Vizați rezultate echitabile pentru diferite grupuri.

- **Conștientizarea dezvoltatorilor**

Asigurați-vă că dezvoltatorii sunt conștienți de potențialele prejudecăți, în special legate de, de exemplu, rasă, culoare, descendență, sex, vârstă, limbă, religie, opinie politică, origine națională, etnică, socială, statut economic, stare de naștere, handicap

- **Comunicare și răspuns**

Stabiliți protocoale clare pentru a comunica, raporta și răspunde prejudecăților, impactului negativ sau compromisurilor.

- **Gestionarea datelor**

Luați în considerare sistemele de stocare a datelor pentru a preveni părtinirea datelor în silozuri.

- **Gestionarea riscurilor**

Elaborați măsuri de tipul "Ce se întâmplă dacă" pentru a face față rezultatelor neașteptate.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Faza de proiectare

Principii etice de luat în considerare



Explicabilitate / Transparență

- Asigurați o documentație clară și explicații ușor de utilizat ale deciziilor AI.
- Asigurați transparența surselor de date și a proceselor algoritmice.

Beneficentă

Implicarea părților interesate în procesul de dezvoltare pentru a lua în considerare diverse aspecte etice.

Autonomie

Implementați mecanisme de supraveghere umană pentru a menține valorile umane în soluțiile AI.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Faza de proiectare Părțile interesate



- **Arhitecți software:** pun bazele tehnice ale soluției AI, luând în considerare atât nevoile actuale, cât și pe cele viitoare.
- **Designeri pentru experiența utilizatorului (UX):** esențială pentru crearea de interfețe care sporesc adoptarea și satisfacția utilizatorilor și pentru a se asigura că utilizatorii pot interacționa eficient cu sistemul AI.
- **Experți în etică:** abordează problemele etice.
- **Profesioniști AI:** colaborarea cu designerii UX pentru a integra componente AI fără probleme.
- **Funcții juridice și etice:** reducerea riscurilor juridice prin abordarea cerințelor de conformitate încă din faza de proiectare și ajutor în conceperea de soluții care respectă reglementările privind protecția datelor și confidențialitatea.
- **Oameni și comunități afectate de modelul IA:** abordare participativă a proiectării.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Colectarea, înțelegerea și pregătirea datelor



După stabilirea obiectivului și a planului de atingere a acestuia, următorul pas constă în lucrul cu datele relevante.

- Adunați toate datele necesare pentru proiect.
- Descrieți, explorați, verificați, selectați, curățați, construiți, integrați și formatați datele în categorii și concepte adecvate scopului urmărit.
- Abordați și rezolvați problemele legate de datele lipsă pentru a asigura exhaustivitatea.
- Recunoașteți și înțelegeți toate ipotezele care stau la baza datelor existente.
- Clarificați modul în care seturile de date vor fi utilizate pentru luarea deciziilor.
- Menținerea unei calități ridicate a datelor, în special în cazurile de informații neclare sau lipsă.

Aceasta este o etapă consumatoare de timp în care se construiește fundația soluției funcționale. Fiabilitatea și performanța soluției sunt direct corelate cu calitatea datelor de bază.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Colectarea, înțelegerea și pregătirea datelor

Principii etice de luat în considerare



Non-maleficență 1/2

- **Integritatea și calitatea datelor care cauzează prejudecăți**

- Abordarea potențialelor prejudecăți care favorizează anumite categorii demografice. Asigurați acuratețea atunci când combinați seturi de date (de exemplu, persoanele cu același nume nu ar trebui să fie amestecate)
- Efectuați verificări amănunțite pentru a evalua și verifica acuratețea datelor, asigurându-vă că acestea nu perpetuează prejudecăți nedrepte.
- Evitați să folosiți date izolate bazate pe date demografice specifice, care pot spori prejudecățile.
- Implementați un depozit de date unificat, cum ar fi un depozit de date, pentru a minimiza părtinirea



(notă: costisitor). Prejudecățile pot fi un factor relevant în soluția AI? De exemplu, asigurarea auto influențată de tipurile de accidente și de caracteristicile demografice implicate. Constrângeri juridice: Sexul nu poate determina tarifele de asigurare auto, conform Curții Europene de Justiție, 2012.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Colectarea, înțelegerea și pregătirea datelor Principii



Non-maleficență 2/2

- **Confidențialitatea și confidențialitatea datelor.**
 - Abordarea riscurilor tipice legate de confidențialitatea și protecția datelor în timpul colectării datelor.
 - Asigurați-vă că sistemele AI respectă confidențialitatea datelor pe tot parcursul ciclului de viață, aderând la reglementări precum GDPR.
- **Reziliența la atac și securitate.**
 - Protejați seturile de date de atacuri și abordați potențialele riscuri de securitate pentru a menține o securitate robustă a datelor.

HLEG (2019) / De Silva & Alahakoon (2022) / Rochel et. AI (2023)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Colectarea, înțelegerea și pregătirea datelor

Principii



- **Oamenii de știință din domeniul datelor** găsesc modele și tendințe în date pentru a le transforma în cunoștințe.
- **ingineri AI:** responsabili pentru evaluarea și verificarea datelor.
- **Analiztii de date** contribuie la înțelegerea caracteristicilor datelor, ajutând la fundamentarea strategiilor de pregătire a datelor.
- **ingineri de date:** responsabili pentru pregătirea datelor și asigurarea calității acestora.
- **Responsabilul cu protecția datelor (DPO):** are responsabilitatea de a se asigura că organizația sa respectă normele atunci când prelucrează datele cu caracter personal ale angajaților, clienților, furnizorilor sau ale altor persoane (denumite "persoane vizate").

CU1 | AI ethics - a practical approach

Eticieni: abordează considerații etice în gestionarea și selectarea datelor.

(Nileș, 2024) (Prinț, 2024) (Săvulescu, 2024) (Comisia Europeană pentru Protecția Datelor, 2024)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Dezvoltarea și formarea modelelor



În această fază, se dezvoltă un model AI pentru a aborda problema definită. Aceasta include mai multe etape critice.

- Selectați și configurați modelul algoritmic și instrumentele, definind în prealabil utilizarea acestora.
- Enumerați și evaluați implicațiile instrumentelor alese, luând în considerare în special potențialele dezavantaje și compromisuri.
- Decideți ce elemente ale modelului trebuie să fie evidențiate și care pot deveni secundare sau chiar ascunse.
- Participați la mai multe runde de rafinare în timpul procesului de formare, îmbunătățind continuu modelul.
- Selectarea instrumentelor implică adesea echilibrarea unor interese contradictorii. De exemplu, trebuie să se decidă dacă este preferabil să se riște ca e-mailurile spam să ajungă în cutia poștală de primire sau ca e-mailurile valide să ajungă în dosarul spam.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Dezvoltarea și formarea modelelor

Principii etice de luat în considerare



Explicabilitate

- În timp ce ne concentrăm asupra aspectelor tehnice, este esențial să verificăm explicabilitatea și dacă există prejudecăți în cadrul sistemului de inteligență artificială, oferind o ocazie excelentă de intervenție.
- Mențineți transparența pe tot parcursul procesului de dezvoltare pentru a vă asigura că părțile interesate îl pot înțelege. Acest lucru implică documentarea clară a surselor de date, a modelelor alese și a raționamentelor din spatele acestor decizii.

Justiție

- Recunoașteți și raportați în mod transparent orice compromisuri între eficiență și alte cerințe.
- Stabilirea unui proces de recunoaștere și evaluare a impactului acestor compromisuri.
- Inginerii AI trebuie să își justifice deciziile, luând în considerare potențialele efecte negative asupra persoanelor afectate.

Morley (2019) Prem (2023)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Dezvoltarea și formarea modelelor

Părțile interesate



- **Cercetătorii AI/ML** transformă definiția problemei într-un model prototipic de AI.
- **Inginerii AI** au experiența necesară pentru a selecta cele mai potrivite instrumente algoritmice pentru anumite scopuri.
- **Oamenii de știință din domeniul datelor** sunt esențiali pentru dezvoltarea modelelor de învățare automată și optimizarea acestora.
- **Experții în etică** sunt esențiali pentru integrarea principiilor etice în procesul de dezvoltare, pentru a evita consecințele neintenționate.
- **Utilizatorii de afaceri** se asigură că modelele îndeplinesc cerințele și așteptările de afaceri.

(D. De Silva & D. Alahakoon, 2022), (J. Rochel & F. Évequoz, 2020)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Evaluarea și testarea modelului



Odată ce modelul a fost dezvoltat și instruit, performanța acestuia trebuie testată și evaluată:

- Evaluați modelul în raport cu obiectivele și criteriile de succes predefinite.
- Explicarea clară a ipotezelor și standardelor pe care se bazează seturile de date.
- Efectuați teste folosind date sau etape nevăzute ale procesului pentru a asigura robustețea.
- Dacă performanța este nesatisfăcătoare, luați în considerare modificarea parametrilor modelului, a arhitecturii sau chiar a seturilor de date din spatele modelului de date.
- Orice deficiențe ar trebui discutate deschis și onest cu liderul de proiect și cu alte părți interesate, deoarece acestea ar putea avea un impact semnificativ asupra persoanelor.
- Conveniți dacă să continuați cu mai multe iterații sau să treceți la faza de implementare.

Această etapă este crucială pentru asigurarea fiabilității și eficienței modelului înainte ca acesta să intre în producție.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Evaluarea și testarea modelului

Principii etice de luat în considerare



Justiție / Beneficență / Non-maleficență

- **Audituri și măsurători etice:**

- Asigurați respectarea principiilor etice, cum ar fi dreptatea, beneficența și non-maleficența, prin audituri și măsurători concepute pentru a putea fi verificate.

- **Securitate și reziliență:**

- Testați reziliența la atacuri și stabiliți un plan de rezervă pentru a asigura securitatea sistemului.

- **Abordarea impacturilor negative și a riscurilor posibile:**

- Implementați strategii pentru minimizarea și raportarea problemelor negative.
- Stabilirea unui mecanism de despăgubire pentru a aborda potențialele prejudicii în cazul în care lucrurile merg prost.
- Formularea factorilor de evaluare a riscurilor cu accent pe confidențialitate, implicații sociale și prejudecăți.

- **Explicabilitate:**

- Explicați clar atât procesele tehnice, cât și deciziile umane implicate în dezvoltarea modelului.

Fiți conștienți de faptul că inginerii supuși presiunii de a finaliza soluțiile pentru implementare se pot confrunta cu un risc crescut de a trece cu vederea aspecte critice.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Evaluarea și testarea modelului

Părțile interesate



- **Inginerii AI** interpretează rezultatele după modelare.
- **Echipele de asigurare a calității (QA) se** asigură că modelele îndeplinesc standardele de calitate și nu conțin erori sau consecințe neintenționate.
- **Experții în etică** asigură utilizarea etică a modelelor AI.
- **Susținătorii/reprezentanții utilizatorilor** se asigură că modelele AI corespund așteptărilor și nevoilor utilizatorilor.
- **Echipele juridice și de conformitate** reduc riscurile juridice prin asigurarea conformității cu reglementările relevante.

(J. Rochel & F. Évequoz, 2020) / (Moore, 2023)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Implementare



Odată ce soluția AI a fost finalizată în mod satisfăcător, aceasta este implementată într-un mediu de producție pentru a aborda probleme din lumea reală. Complexitatea acestei faze depinde de soluția în sine.

- Inițial, soluția este de obicei prezentată unui grup mai restrâns de experți și utilizatori.
- Fie integrați modelul în sistemele existente, fie creați o aplicație sau un serviciu care utilizează modelul sau valorificați informațiile într-un context offline, cum ar fi un raport către conducere.
- Stabiliți o buclă de feedback pentru monitorizarea performanței și a rezultatelor, cu iterații ulterioare, dacă este necesar.
- Punerea în aplicare a strategiilor de gestionare a riscurilor pentru a aborda problemele potențiale în timpul desfășurării.
- Planificați monitorizarea și întreținerea continuă în faza operațională.
- Analizați întregul proiect și elaborați un raport final care să documenteze procesul și rezultatele.

Această abordare structurată asigură integrarea eficientă a soluției AI și capacitatea acesteia de a funcționa conform așteptărilor în scenarii reale.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Implementare

Principii etice de luat în considerare



Autonomie

- **Agenția umană**

- Asigurați-vă că utilizatorii sunt conștienți atunci când deciziile, conținutul, sfaturile sau rezultatele sunt generate de o AI. Evidențiați interacțiunea cu soluția AI pentru a preveni încrederea excesivă.

- **Supravegherea umană**

- Garantarea existenței unei supravegheri umane pentru a interveni atunci când este necesar, asigurându-se că deciziile pot fi revizuite sau modificate de către oameni.

Justiție

- Implementați măsuri puternice de audit pentru a evalua periodic performanța sistemului și respectarea normelor etice.
- Să stabilească mecanisme prin care utilizatorii să poată contesta deciziile sau să depună plângeri, asigurându-se că există protecție pentru avertizori.

Unesco (2023) / Morley (2019)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Implementare

Principii etice de luat în considerare



Explicabilitate

- Subliniați importanța explicării atât a proceselor tehnice, cât și a deciziilor umane aferente pentru utilizatori.
- Abordarea provocării de a simplifica comportamentele umane complexe în instrumente care sunt ușor de înțeles, evitând jargonul tehnic -> acest lucru crește încrederea și acceptarea sistemului
- Mențineți transparența cu privire la funcționalitățile și limitările AI.

Evaluarea impactului pe termen lung

- Evaluați în această etapă impactul pe termen lung asupra societății și mediului.
- Luați în considerare îmbunătățirile necesare pentru a vă asigura că sistemul AI contribuie pozitiv în timp.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Implementare Părțile interesate



- **Inginerii AI/ML** transformă modelul prototipic de AI într-un serviciu sau o soluție implementată care este accesibilă tuturor părților interesate și utilizatorilor finali.
- **Experții în securitate** protejează sistemul implementat împotriva riscurilor de securitate cibernetică.
- **utilizatorii finali** pentru a identifica orice probleme sau limitări care nu ar fi fost evidente în timpul testării în medii controlate.
- **Susținătorii/reprezentanții utilizatorilor** colectează feedback-ul utilizatorilor cu privire la utilitatea și performanța soluției AI implementate și se asigură că soluția implementată răspunde așteptărilor și nevoilor utilizatorilor.

(D. De Silva & D. Alahakoon, 2022) / (Junklewitz et al., 2023)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Funcționare și monitorizare



După implementare, soluția AI necesită întreținere continuă, actualizări și evaluare continuă.

- Asigurați-vă că modelul continuă să funcționeze conform așteptărilor prin monitorizarea periodică a eficacității sale.
- Monitorizați în permanență activitatea utilizatorilor finali pentru a obține informații operaționale și a detecta eventualele probleme.
- Actualizați periodic modelul cu cele mai noi date pentru a-i menține acuratețea și relevanța.
- Rafinarea modelului pe baza feedback-ului utilizatorilor pentru a îmbunătăți funcționalitatea și performanța.
- Actualizările iterative sunt adesea necesare și ar trebui privite ca o parte normală a ciclului de viață al IA, nu ca un eșec.
- Monitorizați randamentul investițiilor și alți parametri relevanți pe baza obiectivelor stabilite inițial.

Această fază este esențială pentru a se asigura că soluția AI rămâne eficientă și continuă să răspundă nevoilor utilizatorilor săi.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Funcționare și monitorizare

Principii etice de luat în considerare



Monitorizarea principiilor etice

- Revedeți și evaluați periodic respectarea principiilor etice de către sistemul de inteligență artificială, luând în considerare posibilele modificări ale seturilor de date și ale structurilor modelului.

Continuarea angajamentului părților interesate

- Menținerea unui angajament activ față de utilizatori, organizații și societatea în general pentru a colecta feedback continuu.
- Stabilirea unor mecanisme structurate pentru a asigura colectarea periodică și sistematică de informații și răspunsuri.

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Funcționare și monitorizare Părțile interesate



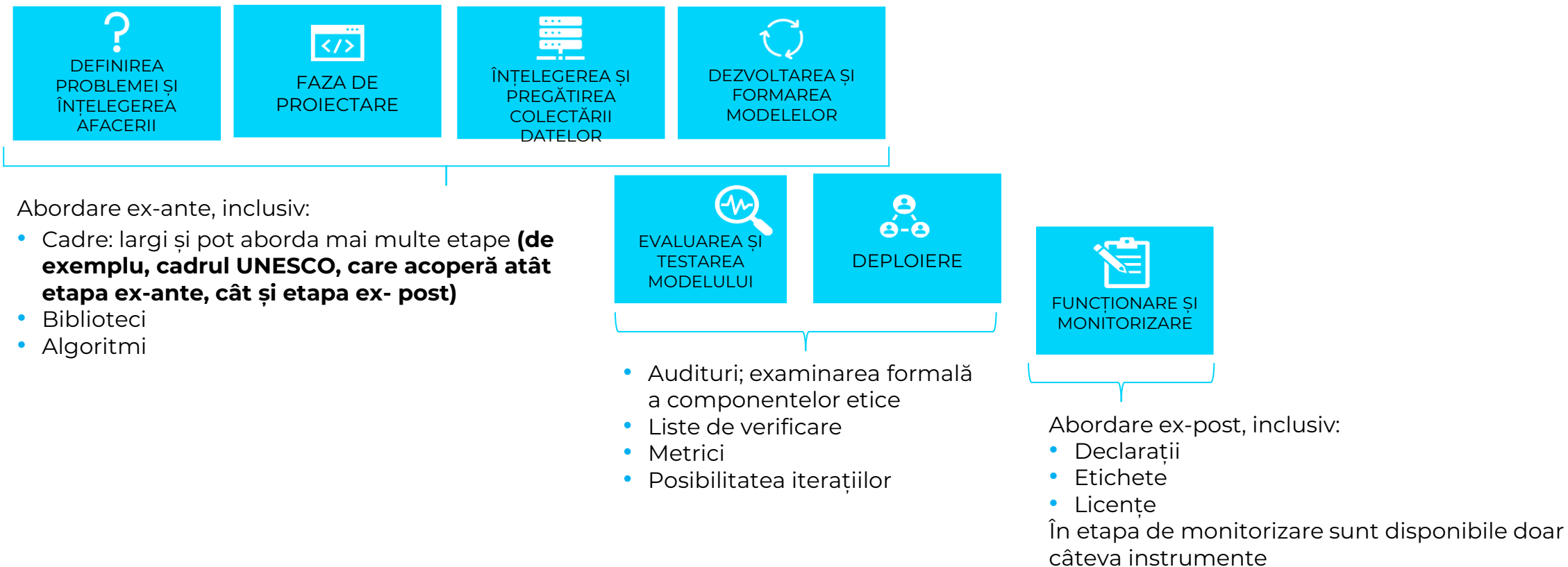
- **DevOps și echipele de operațiuni IT** monitorizează performanța sistemului și asigură disponibilitatea continuă.
- **Certificatorii de siguranță** evaluează securitatea soluției AI dezvoltate.
- **Grupuri de experți, comitete directoare sau organisme de reglementare** analizează proiectul din punct de vedere tehnic și etic.

(D. De Silva & D. Alahakoon, 2022) / (Immersant Data Solutions, 2023) / (Miller, 2022)

Principii etice și părți interesate în ciclul de viață al dezvoltării IA

Ciclul de viață al dezvoltării AI Instrumente

Instrumentele disponibile pentru considerentele etice variază foarte mult în funcție de stadiul de dezvoltare a IA. Multe abordări pot fi valabile în toate etapele, dar unele numai în anumite etape de dezvoltare. Categorizarea de mai jos este o imagine aproximativă a modului în care instrumentele disponibile pot fi împărțite între etapele de proiectare.



VĂ MULTUMESC



Universitat
de les Illes Balears



helixconnect



Cofinanțat de
Uniunea Europeană

Finanțat de Uniunea Europeană. Punctele de vedere și opiniile exprimate aparțin, însă, exclusiv autorului (autorilor) și nu reflectă neapărat punctele de vedere și opiniile Uniunii Europene sau ale Agenției Executive Europene pentru Educație și Cultură (EACEA). Nici Uniunea Europeană și nici EACEA nu pot fi considerate răspunzătoare pentru acestea.

2022-1-ES01-KA220-HED-000085257