Elaborarea cadrului strategic național în domeniul inteligenței artificiale

Analiza reglementărilor pentru domeniul inteligenței artificiale

21 Decembrie 2021

Adrian Groza
George Bara
Cristina Belba
Aurelian Ionescu
Marian Iurian
Camelia Lemnaru
Luciana Morogan
Eugen Popescu

Proiect: Proiect selectat în cadrul Programului Operațional Capacitate Administrativă cofinanțat de Uniunea Europeană, din Fondul Social European
Axa Prioritară 1: Administrație publică și sistem judiciar eficiente
Obiectivul specific 1.1: Dezvoltarea și introducerea de sisteme și standarde comune în administrația publică ce optimizează procesele decizionale orientate către cetățeni și mediul de afaceri, în concordanță cu SCAP.
Titlul proiectului: „Cadru strategic pentru adoptarea și utilizarea de tehnologii inovative în administrația publică 2021-2027 – soluții pentru eficientizarea activității”
COD: SIPOCA 704
Beneficiar: AUTORITATEA PENTRU DIGITALIZAREA ROMÂNIEI
Partener: UNIVERSITATEA TEHNICĂ DIN CLUJ - NAPOCA

Cuprins

Rezumat

Raportul analizează politicile emergente pentru reglementarea inteligenței artificiale (IA) și interacțiunea cu reglementările existente în domenii conexe. Raportul are două părți. Prima parte vizează cadrul de reglementare al Comisiei Europene pentru inteligența artificială, integrarea acestuia în politicile emergente pentru dezvoltarea IA etice și interacțiunea cu reglementările din alte sectoare. A doua parte prezintă o posibilă arhitectură instituțională pentru reglementarea IA în România. Viziunea presupune *o autoritate națională competentă* pentru monitorizarea și reglementarea IA etice și un model descentralizat în care organisme de evaluare a conformității (e.g. centre de audit - private sau publice) au competențe de monitorizare, verificare și certificare în diferite subdomenii și tehnologii ale inteligenței artificiale.

Capitolul 1 prezintă strategia Comisiei Europene (CE) pentru reglementarea IA. CE plasează agentul uman în centrul dezvoltării IA și urmărește dezvoltarea unei inteligențe artificiale etice. Scopul reglementării inteligenței artificiale în UE este de a îmbunătăți *funcționarea pieței interne* prin stabilirea unui *cadru juridic uniform* care să susțină dezvoltarea, comercializarea și utilizarea IA, în conformitate cu valoriile UE. CE propune o abordare bazată pe clasificarea aplicațiilor de inteligență artificială în patru niveluri diferite de risc: "risc inacceptabil", "risc ridicat", "risc limitat", respectiv "risc minim". Capitolul descrie aceste grupe de risc și exemplifică aplicații IA conform posibilei încadrări a acestora în diferite niveluri de risc.

Capitolul 2 prezintă instituții create în diferite state pentru monitorizarea inteligenței artificiale sau pentru implementarea strategiilor naționale pe IA. Abordările naționale sunt diferite în funcție de percepția riscurilor și beneficiilor asociate cu IA sau în funcție de domeniu de aplicabilitate.

Capitolul 3 introduce aspectele etice al IA ca bază a reglementărilor, în contextul în care UE dorește să fie liderul în dezvoltarea de IA responsabil. Cadrul de reglementare propus de CE pune pe prima poziție aspectele etice precum transparența algoritmilor, reducerea discriminării, supravegherea umană . După prezentarea acestor principii, capitolul enumeră un set de ghiduri care se referă IA etică, IA responsabilă sau IA de încredere. Aceste ghiduri pot fi puncte de pornire pentru operaționalizarea inteligenței artificiale etice și în implementarea procedurilor pentru evaluarea conformității și certificarea aplicațiilor IA etice. De asemenea, sunt prezentate pe scurt politicile pentru IA etică a Google, Axon, DeepMing, Microsoft, sau Amazon, precum și unelte de IA dezvoltate recent pentru a sprijini IA etică, inclusiv pe linia Explainable AI (XAI).

Capitolul 4 exemplifică o serie de aplicații IA în diferite sectoare de activitate și face referiri la posibila clasificare a acestor aplicații din perspectiva grupelor de risc propuse de AIA. Sunt exemplificate o parte din reglementările în vigoare, în contextul în care AIA este un cadru de reglementare pentru inteligența artificială care se bazează pe standarde și pe utilizarea pe cât posibil a experienței și reglementărilor din sectoarele specifice. Cum IA va afecta toate sectoarele de activitate, capitolul argumentează că prioritizarea sectoarelor cu scopul de a facilita dezvoltarea IA pentru un domeniu specific ar fi o decizie bazată pe multe necunoscute. De aceea, o abordare în care "pătrunderea sistemelor IA este sprijinită echidistant în toate sectoarele" poate fi o soluție corectă. Principalul criteriu pentru decizii curente legate de finanțarea proiectelor IA ar putea fi *valoarea adăugată adusă de IA în domeniul sau scenariul respectiv*. La nivel UE, trei domenii în care IA aduce cea mai mare valoare adăugată ar putea fi: (i) canale de distribuție, (ii) mobilitate și orașe inteligente, (iii) sănatate.

Capitolul 5 argumentează în favoarea reglementării campaniilor de dezinformare care se bazează din ce în ce mai mult pe unelte IA. Dezinformarea în România reprezintă o vulnerabilitate și afectează direct funcționarea societății la scară largă, cu riscul de a afecta grav sănătatea sau democrația, așa cum o dovedește exemplul social recent de reticență la vaccinarea anti-Covid. Dezvoltarea prioritară de unelte de inteligență artificială pentru combaterea dezinformării sau pentru reducerea ignoranței poate aduce beneficii semnificative la nivelul societății. Cadrul de reglementare pentru combaterea dezinformării rămâne o provocare datorită dificultăților de a trage linie clară între

campanii de dezinformare și libertatea de expresie. Capitolul prezintă inițiativele de reglementare și măsuri specifice referitoare la dezinformare în state precum Marea Britanie, Germania, Singapore, Cehia, Bulgaria, Moldova, Suedia, Rusia. Abordarea industriei (i.e. platformele sociale) pe linia dezinformării este exemplificată prin prezentarea standardelor Facebook, a politicilor Google și YouTube sau a abordării TikTok. Sunt discutate aspecte legate de reglementarea generării de conținut sintetic video sau audio (i.e. deepfake), atât din pespectiva Comisiei Europene, cât și din cea a inițiativelor Facebook sau Google. De asemenea, sunt exemplificate o serie de unelte IA folosite pentru combaterea dezinformării. Capitolul se încheie cu o schiță de recomandări pentru combaterea dezinformării în România prin înființarea unui birou specializat în cadrul Consililului Național al Audiovizualului, în colaborare cu Autoritatea de Reglementare a Inteligenței Artificiale.

Capitolul 6 schițează funcțiile de bază, atribuțiile și drepturile generale ale unei autorități pentru reglementarea inteligenței artificiale. Conform "Artificial Intelligence Act" al CE, se dorește inființarea unei astfel de autorități in fiecare stat membru începând cu 2023. Capitolul identifică instituții cu care această autoritate va coopera pentru monitorizarea siguranței aplicațiilor IA și a asigurării dezvoltării IA etice. De asemenea, autoritatea va desemna organisme de evaluare a conformității aplicațiilor IA la nivel național.

Capitolul 7 detaliază aspecte legate de evaluarea conformității aplicațiilor bazate pe IA. Cadrul pentru reglementarea inteligenței artificiale va avea în centru conceptul de certificare. Această certificare se va realiza de către organisme de evaluare a conformității desemnate de ARIA. Actul pentru Inteligență Artificială vine în completarea legislației sectoriale, cerințele pentru sistemele de IA prevăzute în propunere fiind verificate ca parte a procedurilor existente de evaluare a conformității pe baza legislației relevante NCL (Noul Cadru Legislativ). Operaționalizarea lui se bazează pe implementarea și respectarea de standarde. Sunt prezentate o parte din activitățile de standardizare care vizează aplicații de inteligență artificială. De asemenea, capitolul descrie câteva din metodele utilizate în auditarea sau asigurarea calității sistemelor dezvoltate cu tehnologii din domeniul inteligenței artificiale. Standardele emergente și metodele de auditare a calității prezentate aici pot constitui un punct de plecare pentru instituirea unor astfel de organisme de evaluare a conformității și de certificare - atât publice cât și private - la nivel național.

Capitolul 8 introduce un instrument important pentru dezvoltarea IA in România, și anume, spațiile de testare în materie de reglementare a IA. Spațiile de testare în materie de reglementare a IA (i.e. "regulatory sandboxes") creează un mediu controlat pentru testarea tehnologiilor inovatoare pentru o perioadă limitată de timp, pe baza unui plan de testare convenit cu autoritățile competente. Este necesar un cadru pentru reglementarea funcționalității acestor spații. În plus, viteza cu care se dezvoltă tehnologiile IA ridică în mod recurent provocări legate de identificarea celui mai bun cadru de reglementare. Apare astfel necesitatea de a experimenta eficacitatea unor reglementări, cu scopul de a identifica cele mai bune reguli. Acesta este unul din rolurile spațiilor de testare în materie de reglementare a IA. Capitolul exemplifică astfel de spații precum și tipurile de clauze de experimentare care apar în aceste medii controlate.

Acest document însoțește Raportul de analiză a abordării europene si mapare a inițiativelor din domeniul inteligenței artificiale de la nivel internațional. Aspecte din acest document vor fi cuprinse în Cadrul strategic național pentru inteligență artificială care va fi finalizat în august 2022.

Reglementarea și Monitorizarea Inteligenței Artificiale

1. Cadrul de reglementare al UE pentru IA

Rezumat. Prezentăm strategia Comisiei Europene (CE) pentru reglementarea inteligenței artificiale (IA). CE plasează agentul uman în centrul dezvoltării IA și urmărește dezvoltarea unei inteligențe artificiale etice. Scopul reglementării inteligenței artificiale în UE este de a îmbunătăți *funcționarea pieței interne* prin stabilirea unui *cadru juridic uniform* care să susțină dezvoltarea, comercializarea și utilizarea IA, în conformitate cu valoriile UE. CE propune o abordare bazată pe clasificarea aplicațiilor de inteligență artificială în patru niveluri diferite de risc: "risc inacceptabil", "risc ridicat", "risc limitat", respectiv "risc minim". Capitolul descrie aceste grupe de risc și exemplifică aplicații IA conform posibilei încadrări a acestora în diferite niveluri de risc.

Strategia UE plasează agentul uman în centrul dezvoltării inteligenței artificiale și urmărește dezvoltarea unei inteligențe artificiale etice [7] pe baza a trei piloni: (i) consolidarea investițiilor publice și private în IA, (ii) pregătirea pentru schimbări socio-economice, (iii) asigurarea unui cadru etic și juridic adecvat [16]. Reglementarea inteligenței artificiale reprezintă deci una din cele trei dimensiuni în strategia UE pentru dezvoltarea IA.

Scopul reglementării inteligenței artificiale în UE este de a îmbunătăți *funcționarea pieței interne* prin stabilirea unui *cadru juridic uniform* care să susțină dezvoltarea, comercializarea și utilizarea IA, în conformitate cu valoriile UE. Conform Propunerii pentru un Regulament al Parlamentului European și al Consiliului de stabilire a unor norme armonizate privind inteligența artificială (Legea privind Inteligența Artificială) și de modificare a anumitor acte legislative ale Uniunii (Propunerea de Regulament IA) din 21.04.2021 [28], cadrul de reglementări privind inteligența artificială prezintă următoarele obiective specifice:

Obiective vizate de UE prin cadrul de reglementare pentru IA

1. Asigurarea faptului că sistemele de IA introduse pe piața Uniunii și utilizate sunt sigure și respectă legislația existentă privind drepturile fundamentale și valorile Uniunii;
2. Asigurarea securității juridice pentru a facilita investițiile și inovarea în domeniul IA;
3. Consolidarea guvernanței și asigurarea efectivă a respectării legislației existente privind drepturile fundamentale și a cerințelor de siguranță aplicabile sistemelor de IA;
4. Facilitarea dezvoltării unei piețe unice pentru sisteme de IA legale, sigure și de încredere, precum și prevenirea fragmentării pieței.

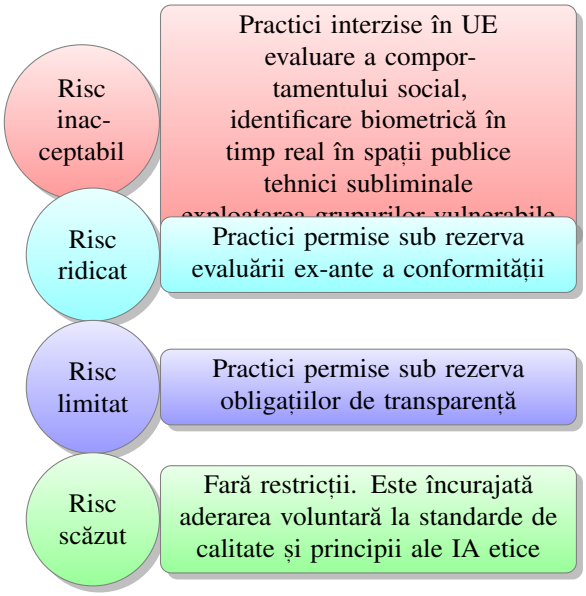

Figura 1.1: Abordarea CE bazată pe niveluri de risc

Propunerea de regulament urmărește mai multe dimensiuni, printre care și asigurarea unui nivel ridicat de protecție a sănătății, a siguranței și a drepturilor fundamentale, precum și asigurarea liberei circulații transfrontaliere a bunurilor și serviciilor bazate pe IA. Cadrul de reglementare propus [28] urmărește ca sistemele de IA utilizate pe teritoriul UE să fie sigure, transparente, etice, imparțiale și să fie sub control uman. Omul este plasat în centru, sistemele IA fiind unelte care sprijină activitatea agenților umani.

Pentru definirea aplicațiilor care vor intra sub incidența reglementărilor specifice inteligenței artificiale, CE utilizează o definiție largă a IA, în care sistemele considerate ca IA sunt împărțite în trei categorii [28].

Definirea sistemelor IA:

1. Abordări bazate pe învățare automată;
2. Abordări bazate pe logică și cunoaștere, inclusiv reprezentarea cunoștințelor, programare inductivă (logică), baze de cunoștințe, motoare inductive și deductive, sisteme de raționament (simbolic) și de expertiză;
3. Abordări statistice: estimare Bayesiană, metode de căutare și optimizare.

R O confuzie recentă dar recurentă legată de inteligența artificială este reducerea IA la învățarea automată (i.e. machine learning). Învățarea automată este un un subdomeniu al IA. Pentru perspective detaliate a ceea ce este inclus în inteligența artificială se pot consulta definițiile IA din [92].

Pentru reglementarea inteligenței artificiale, Comisia Europeană propune o abordare bazată pe risc [23] [28][1]. Se face distincția între patru niveluri de risc: “risc inacceptabil”, “risc ridicat”, “risc limitat”, respectiv “risc minim” (Figura 1.1).

1.1 Aplicații și practici IA interzise

[1] Un rezumat al propunerii UE pentru reglementarea IA [28] poate fi consultat la [30].

Risc inacceptabil. Practicile considerate a fi o amenințare clară la adresa siguranței, a mijloacelor de subzistență și a drepturilor cetățenilor UE vor fi interzise. Sunt enumerate patru categorii:

1. Evaluarea comportamentului social care este făcută prin intermediul sistemelor de IA în scopuri generale de către autoritățile publice
2. Utilizarea de tehnici subliminale
3. Explotarea vulnerabilităților unor persoane sau grupuri de persoane
4. Identificare biometrică în spații publice

R Se dorește protejarea drepturilor *persoanelor fizice*, ceea ce este o extensie față de reglementările existente care vizează categorii specifice precum *data subjects* sau *clienți*.

R Primele trei practici sunt interzise pentru a fi plasate pe piață sau utilizate, iar cea de-a patra (i.e. identificare biometrică în timp real în spații *fizice* accesibile publicului) are prevăzute excepții. Statele membre pot decide intern asupra acestor excepții.

R Reglementarea vizează deci atât furnizorii sau dezvoltatorii de soluții IA, cât și potențialii utilizatori.

R Reglementările pentru practicile interzise se aplică atât actorilor publici cât și celor din mediu privat.

R Reglementarea nu se aplică activității de cercetare, cu două mențiuni: (i) cercetarea urmează standarde etice, respectiv (ii) cercetarea nu este utilizată prin interacțiune om-mașină [42].

R Fiind un regulament, se aplică doctrina *efectului direct* prin care se poate invoca în mod direct normele europene în fața instanțelor naționale sau europene (cnf. Art. 288 din Tratatul prinvind funcționarea UE).

R Pentru a avea efect preventiv, amenzile sunt semnificative; până la 30 milioane sau 6% din cifra de afaceri anuală. Conform propunerii de regulament, nerespectarea interdicției privind practicile în domeniul inteligenței artificiale menționate la articolul 5 și neconformitatea sistemului de IA cu cerințele prevăzute la articolul 10, sunt încălcări ce fac obiectul unor amenzi administrative de până la 30 000 000 EUR sau, în cazul în care autorul infracțiunii este o întreprindere, de până la 6% din cifra sa de afaceri mondială totală anuală pentru exercițiul financiar precedent, luându-se în considerare valoarea cea mai mare. (statele membre pot decide dacă și în ce măsură aplică aceste amenzi).

R Persoanele fizice lezate prin utilizarea uneia din practicile interzise ale IA nu primesc drept explicit de a a face plângeri sau despăgubire. Autoritățile responsabile pentru monitorizarea pieței (e.g. autoritățile naționale pentru reglementarea inteligenței artificiale) pot iniția respectivele plângeri. Autoritățile pot iniția acțiuni pentru a preveni plasarea iminentă unei astfel de aplicații în Zona Economică Europeană [42].

Manipularea persoanelor prin tehnici subliminale cu scopul de a afecta comportamentul persoanelor și care poate cauza provoca efecte fizice sau psihologice. Drepturile fundamentele care sunt amenințate de astfel de aplicații sunt:

- Dreptul la demnitate (Art. 1 Carta drepturilor fundamentale a UE)
- Dreptul la integritate fizică și mentală (Art. 6 1 Carta drepturilor fundamentale a UE)
- Libertatea de gândire, conștiință și religie (Art. 10 1 Carta drepturilor fundamentale a UE)

Pentru a se încadra în această categorie, un sistem IA trebuie să îndeplinească 3 condiții [42]:

1. să afecteze persoanele fără ca acestea să fie conștiente
2. sunt afectate autonomia și abilitatea persoanei de a acționa astfel încât compartamentul acesteia este diferit față de situația în care nu ar fi interacționat cu sistemul IA
3. cauzează sau există riscul de a cauza pagube fizice sau psihologice.

■ **Exemplu 1.1 — Manipularea persoanelor prin tehnici subliminale.** Aplicații de realitatea augmentată în care experiența senzorială este controlată pentru a incita utilizatorul să facă acțiuni periculoase. ■

■ **Exemplu 1.2 — Manipularea persoanelor prin tehnici subliminale.** Asistenți virtuali care dau sfaturi legate de diete periculoase proiectați pentru a crește beneficiile economice ale unor agenți economici. ■

■ **Exemplu 1.3 — Manipularea persoanelor prin tehnici subliminale.** Aplicații IA a căror interfață a fost proiectată utilizând șabloane întunecate (i.e "Dark Patterns"), de exemplu încorporate in jocuri video și care captează excesiv utilizatorii. ■

Exemple de șabloane întunecate. Chris Lewis analizează 27 șabloane de proiectare "motivaționale" care urmăresc atragerea utilizatorilor spre anumite aplicații [61]. Șabloanele de proiectare a aplicațiilor "irezistibile" sunt grupate în 7 categorii: șabloane ludice, șabloane sociale, șabloane pentru interfețe, șabloane de informare, șabloane temporale întunecate, șabloane monetare, șabloane de capital social. Rămâne o discuție în ce măsură astfel de șabloane pot fi interpretate ca instrumente tehnice de manipulare subliminală, conform viziunii AIA.

Exploatarea vulnerabilităților unor categorii vulnerabile specifice în funcție de vârstă sau a dizabilităților fizice sau mentale pentru a afecta în mod semnificativ comportamentul unei persoane care aparține grupului respectiv într-un mod care poate aduce prejudicii fizice sau psihologice; și a provoca riscuri fizice sau psihologice. Suplimentar drepturilor afectate anterior, această categorie de aplicații IA afectează drepturile persoanelor defavorizate (Articolele 24, 25, 26 1 Carta drepturilor fundamentale a UE)

Pentru a se încadra în această categorie, un sistem IA trebuie să îndeplinească cumulativ 3 condiții [42]:

1. Să exploateze o vulnerabilitate unor categorii vulnerabile specifice: imaturitate, dependență lipsa auto-controlului, comportament inclinat spre risc, incapacitate fizică, fragilitate psihică;
2. Să existe intenția de a afecta comportamentul;
3. Comportamentul astfel indus să introducă riscul unei vătămari fizice sau psihice.

■ **Exemplu 1.4 — Exploatarea vulnerabilităților unor unor categorii vulnerabile specifice.** Aplicații proiectate pentru a induce adicție sau comportament compulsiv la copii (e.g. recompense aleatorii, trimiterea de notificații când aplicația nu este utilizată. Vulnerabilitățile exploatate în acest caz sunt lipsa auto-controlului, dependența, fragilitate psihică. ■

■ **Exemplu 1.5 — Exploatarea vulnerabilităților unor categorii vulnerabile specifice.** Un robot medical proiectat să convingă persoanele în vârstă să respecte o anumită rutină zilnică, indiferent de voința persoanei vizate. Vulnerabilitățile exploatate sunt: dependență, fragilitate

psihică, capacitate fizică redusă. ■

■ **Exemplu 1.6 — Exploatarea vulnerabilităților unor categorii vulnerabile specifice.** Un asistent bazat pe senzori (inclusiv analiză video) pentru persoane cu disabilități care pun viața respectivei persoane în pericol. Incapacitatea fizică sau dependența sunt două vulnerabilități vizate. ■

R Sunt două diferențe între aplicațiile IA care manipulează subliminal și cele care exploatează vulnerabilitățile unor categorii vulnerabile specifice. În primul rând, pentru aplicațiile IA cu manipulare prin tehnici subliminale, orice persoană fizică este protejată. În schimb, pentru aplicațiile care exploatează vulnerabilitățile unor categorii vulnerabile specifice, doar membrii respectivului grup sunt protejați. În al doilea rând, dacă tehnicile subliminale sunt ascunse, tehnicile de exploatare a vulnerabilitătilor nu sunt acceptate chiar dacă utilizatorul este conștient de ele.

■ **Exemplu 1.7 — Aplicații IA cu tehnici subliminale care exploatează și vulnerabilitățile unor categorii vulnerabile specifice.** Sistem IA care recunoaște emoțiile unor adolescenți și face recomandări personalizate de piese muzicale depresive celor detectați ca fiind "într-o dispoziție proastă" și care amplifică discomfortul psihic. ■

R O parte din aplicațiile IA interzise pot intra și sub incidența Directive 29 din 2005 (UCPD) cu privire la practicile comerciale incorecte. Această directivă este limitată însă la relațiile economice. În caz de conflict între UCPD și AIA (Artificial Intelligence Act), conform principiului lex specialis, AIA va fi utilizat.

R Aplicațiile de tip manipulare prin tehnici subliminale și cele care exploatează vulnerabilități ale unor categoriilor vulnerabile specifice pot intra și sub incidența DSA (Digital Services Act) dacă sunt furnizate sau utilizate de către platformele online. DSA prevede obligații de gestiune a riscului pentru ca platformele online să respecte constrângerile impuse de reglementările de siguranța și protecție a consumatorului. De exemplu, obligațiile de transparență pentru reclame țintite pot elimina riscul de manipulare prin tehnici subliminale [42].

Evaluare a comportamentului social, care este făcută prin intermediul sistemelor de IA în scopuri generale de către autoritățile publice prin utilizarea unor aplicații care evaluează aspecte legate de încrederea în persoane fizice pe baza comportamentului social și care conduce la (i) tratament nefavorabil în contexte care nu au legătură cu datele colectate pentru calcularea scorului, sau (ii) tratament nefavorabil nejustificat sau disproporționat cu comprtamentul social al persoanei vizate. Drepturile afectate prin evaluare comportamentului social sunt [42]:

- Dreptul la demnitate (Art. 1 Carta drepturilor fundamentale a UE)
- Dreptul la viață privată și protecția datelor personale (Art. 7 și 8 Carta drepturilor fundamentale a UE)
- Dreptul la egalitate și non-descriminare (Art. 20 și 21 Carta drepturilor fundamentale a UE)
- Dreptul la solidaritate - socială, sănatate (Art. 34, 35 și 36 Carta drepturilor fundamentale a UE)
- Dreptul la bună adminstrație (Art. 41 Carta drepturilor fundamentale a UE)

Pentru ca evaluare a comportamentului social să se încadreze în această categorie, sunt necesare 3 condiții [42]:

1. să evalueze aspecte legate de încrederea într-o persoană fizică, pe un interval de timp, pe baza comportamentului social sau caracteristicile de personalitate cunoscute sau prezise ale persoanei vizate
2. să fie calculată de autoritățile publice sau de actori privați la cererea sau în numele unei autorități publice

3. să existe o relație cauză efect între evaluarea comportamentului social și un tratament nefavorabil aplicat disproporționat sau în contexte diferite

■ **Exemplu 1.8 — Evaluarea comportamentului social - date utilizate în alte contexte.** O autoritate fiscală investighează posibile fraude ale unor persoane fizice pe baza unui sistem IA de analiză a datelor care utilizează și date - de exemplu de pe rețelele sociale sau date legate de activitățile zilnice ale persoanei vizate. ■

■ **Exemplu 1.9 — Evaluarea comportamentului social - tratament nejustificat sau disproporționat.** Un serviciu de protecție socială analizează folosind unelte IA de evaluare a riscului daca beneficiarii subvențiilor au comis fraude, cu scopul de a limita consumul de electricitate pe o perioadă de câteva luni. ■

■ **Exemplu 1.10 — Evaluarea comportamentului social - date în alte contexte și tratament disproporționat.** Un sistem IA care identifică copiii cu nevoi sociale, pe baza unor factori nesemnificativi sau irelevanți (e.g. comportamentul părinților în contexte diferite). ■

R Nu sunt interzise aplicațiile IA care evaluează comportamentului social pentru companii/persoane juridice sau orice evaluarea a comportamentului social a entităților private (care nu acționează in numele unei autorități publice).

R Aplicațiile IA care evaluează comportamentul social pot intra și sub incidența Legii nondiscriminării, dar care se aplică doar pentru discriminare pe baza unei liste explicite de caracteristici (e.g. gen, rasă, religie) și doar pentru protecție socială sau servicii publice. AIA vizează orice tratament nefavorabil și nu necesită demonstrarea ca un tratament mai favorabil a fost aplicat membrilor unui grup [42].

R Aplicațiile IA care evaluează comportamentul social pot intra și sub incidența Legii protecției datelor. Principiile din GDPR (e.g. corectitudine, limitarea scopului, minimizarea datelor) se aplică și autorităților publice. AIA suplimentează aceste principii cu o interzicere explicită a evaluării comportamentului social realizată prin încălcarea dreptului la viață privată sau protecția datelor [42].

R AIA poate fi activitat indiferent dacă există sau nu legislație specifică la nivelul statelor membre pentru aplicații IA care evaluează comportamentul social.

Se consideră că sistemele de identificare biometrică prezintă un grad de risc ridicat. Din acest motiv, acestea sunt supuse unor reguli stricte, iar ca primă regulă este interzisă utilizarea lor în timp real în spații accesibile publicului în scopuri de asigurare a respectării legii. Sunt prevăzute și unele excepții, strict definite și reglementate [42].:

1. Urmărirea țintită în cazul unor potențiale victime (e.g. dispariția copiilor)
2. Prevenirea unei amenințări iminente pentru viața si securitatea persoanelor sau în caz de atac terorist
3. Detectarea, localizarea, identificarea și urmărirea în justiție a unor persoane suspecte de infracțiuni cu mandat de arestare european de cel puțin 3 ani.

Aceste utilizări excepționale sunt condiționate de obținerea ex-ante a unei autorizații din partea unui organism judiciar, de acoperirea geografică și de bazele de date folosite pentru efectuarea căutărilor. Acordarea autorizației trebuie să ia în calcul drepturile fundamentale, să prevadă limitări și măsuri de siguranță și sa fie limitate în timp si arie geografică. În caz de urgență autorizarea poate fi și ex-post.

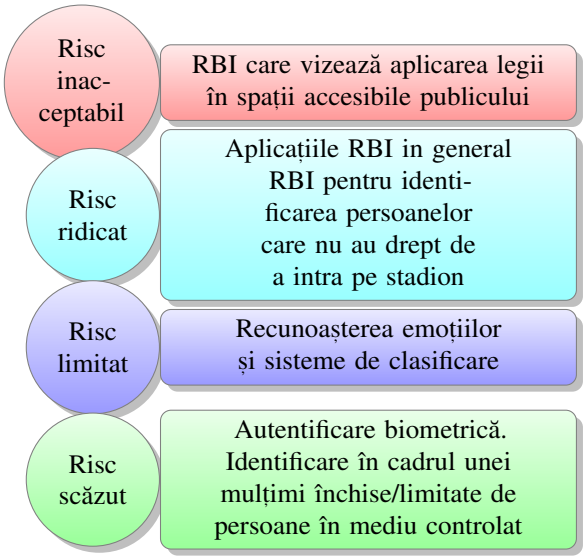

Figura 1.2: Identificarea biometrică la distanță (RBI) poate fi implementată în aplicații care aparțin unor grupe de risc diferite

Încadrarea aplicațiilor IA în această categorie se face pe baza a 5 elemente cumulative [42].:

1. Aplicația să servească aplicării legii (prevenția, investigarea, urmarirea unor infracțiuni);
2. Aplicația identifică biometric un individ într-un grup de persoane;
3. Aplicația operează identificare biometrică la distanță într-un mediu necontrolat;
4. RBI se realizează în timp real;
5. Să opereze în spații accesibile publicului, indiferent dacă acest spațiu este public sau privat.

R Tehnologiile care permit evaluare biometrică pot fi implementate în aplicații IA care aparțin unor grupe de risc diferite (Fig. 1.2)

Evaluarea biometrică în spații publice (RBI). Sunt considerate practici care interferează cu drepturile și libertățile fundamentale datorită sentimentului constant de supraveghere, impactului imediat și potențialului la nivel psihologic de a altera comportamentul natural ale persoanelor care s-ar afla in astfel de spații publice supravegheate. Drepturi afectate de astfel de practici sunt [42].:

- Dreptul la demnitate (Art. 1 Carta drepturilor fundamentale a UE)
- Dreptul la viață privată și protecția datelor personale (Arts. 7 și 8 Carta drepturilor fundamentale a UE)
- Dreptul la egalitate și non-doscriminare (Art. 20 și 21 Carta drepturilor fundamentale a UE)
- Dreptul la libera asociere (Art. 12, Carta drepturilor fundamentale a UE)

■ **Exemplu 1.11 — Identificare biometrică în spații accesibile publicului.** Identificarea feței într-o piață publică prin camere video și verificarea în timp real a acestor fețe cu o bază de date a unei agenții de aplicare a legii. ■

R Identificările care nu au loc in timp real nu intră sub incidența acestui articol (e.g. analiza imaginilor video după producerea unui incident pentru a identifica suspecții).

Nefiind spații accesibile publicului, închisorile sau spațiile de birouri nu intră sub incidența acestui articol din regulament. Fiind în zonă controlată, identificarea biometrică la aeroport este permisă.

■ **Exemplu 1.12 — Identificarea în timp real realizată de entități private în spații publice.** Un club de fotbal folosește recunoaștere facială in timp real lângă stadion pentru identificarea persoanelor care nu a dreptul de a intra pe stadion. Pentru a obține aprobare pe partea de protecție a datelor (deoarece acest scenariu intră în grupa de risc ridicat), pot fi cerute anumite măsuri de siguranță (e.g. sistemul IA de identificare să nu aibă legătură la Internet [42]. ■

R GDPR interzice în principiu utilizarea sistemelor de identificare biometrică, dar nu vizează direct aplicarea legii, aceasta fiind acoperită de Directiva "Law Enforcement" (LED). Articolul 10 din LED permite sisteme de identificare biometrică sub condiția autorizării de către statele membre. În raport cu cadrul de reglementare LED, pentru aplicațiile menționate în explicit în AIA se aplică doctrina lex specialis. Statele membre vor stabili reguli de autorizare în cazurile speciale [42].

Exemple de excepții pentru utilizare de aplicații IA RBI sunt

■ **Exemplu 1.13 — RBI: căutarea potențialelor victime ale unei infracțiuni.** În cazul răpirilor, autoritățile pot utiliza identificarea biometrică la distanță în timp real pentru identificarea victimei sau a infractorului. ■

■ **Exemplu 1.14 — RBI: Atac terorist.** În cazul unor indicatori credibili ai unui atac terorist pe un stadion, RBI poate fi activat în puncte strategice, cu câteva ore înainte de începerea evenimentului. ■

■ **Exemplu 1.15 — RBI: infrațiune gravă.** Există informații credibile conform cărora un suspect de crimă se află într-un anumit tren. RBI poate fi activat pentru o perioadă limitată de timp pentru supravegherea gărilor în care suspectul ar putea fi la un anumit moment de timp. ■

R O lege pentru supraveghere în masă se află din Septembrie 2021 în dezbatere publică în Serbia. În caz de adoptare, Serbia va deveni prima țară europeană în care este permisă monitorizarea biometrică în spații publice.

1.2 Aplicații de IA cu risc ridicat

Aplicațiile de IA cu grad ridicat de risc sunt permise pe piața europeană sub rezerva respectării anumitor cerințe obligatorii și a unei evaluări ex ante a conformității. Sistemele de IA din categoria celor cu risc ridicat sunt cele care pot afecta siguranța utilizatorilor. Fiind vizată siguranța produselor, această categorie intră în primul rând sub incidența reglementărilor sectoriale specifice pentru produse cu grad de risc (e.g. dispozitive medicale, vehicule). Ca urmare, AIA (i) reutilizează definițiile existente legate de siguranța si riscuri, (ii) se bazează pe metodele specifice de evaluare a riscurilor in diferite sectoare (e.g. evaluare riscurilor de către o parte terță); și (iii) se bazează pe instrumentele sectorului respectiv de evaluare a conformității [41].

R Categoria de risc ridicat vizează două categorii: (i) sisteme de IA destinate a fi utilizate drept componente de siguranță ale produselor care fac obiectul unei evaluări ex ante a conformității de către terți; (ii) alte sisteme de IA autonome, cu implicații în principal asupra drepturilor fundamentale.

Pentru a se încadra in aceasta categorie trebuie îndeplinite două condiții [41]:

1. produsul să fie o componentă de siguranță conform legislației în vigoare
2. produsul cu componentă IA se încadrează în a fi evaluat conform legislației sectoriale de către o parte terță;

R Necesitatea evaluării de către o parte terță a produsului este criteriul de încadrare în grupa de risc ridicat

Armonizarea reglementărilor pentru aplicații IA cu risc ridicat este acoperită de două abordări [41]:

1. Noul Cadru Legislativ (New Legislative Framework - NLF) care introduce cerințe de nivel înalt operaționalizate prin standardizare
2. Cadre de reglementare in care autoritățile publice au un rol mai puternic în aprobarea intrării pe piață a unui produs

Armonizarea prin directive de tip Noul Cadru Legislativ (NLF)

1. Dir. 2006/42/EC pentru mașinarii și utilaje
2. Dir. 2009/48/EC pentru siguranța jucăriilor
3. Dir. 2013/53/EU produse recreaționale craft/watercraft
4. Dir. 2014/33/EU pentru lifturi și subcomponente
5. Dir. 2014/34/EU pentru echipamente utilizate in atmosferăcu risc de explozie
6. Dir. 2014/53/EU pentru echipamente radio
7. Dir. 2014/68/EU pentru echipamente de control a presiunii in tuburi de gaz
8. Reg. (EU) 2016/424 pentru instalații de transport cu cabluri
9. Reg. (EU) 2016/425 pentru echipamentele de protecție individuală (e.g. căști)
10. Reg. (EU) 2016/426 pentru dizpozitive pe gaz sau combustibil (e.g. aragaz)
11. Reg. (EU) 2017/745 pentru dispozitivele medicate (e.g. scaner MRI, robot chirurgical)
12. Reg. (EU) 2017/746 pentru dispozitive medicale de diagnostic in vitro (e.g. teste de sarcina, HIV)

Armonizarea prin cadre de reglementare care vizează pre-aprobarea

1. Reg. (EC) 300/2008 pentru securitatea aviației civile (e.g. echipamente de securitate la aeroport)
2. Reg. (EU) 168/2013 pentru vehicule cu două sau trei roți (e.g. mopede)
3. Reg. (EU) 167/2013 pentru vehicule agricole și forestiere
4. Dir. 2014/90/EU pentru dispozitive marine (e.g de navigare)
5. Dir. (EU) 2016/797 pentru interoperabilitatea sistemului feroviar (e.g. componente critice pentru deplasare)
6. Reg. (EU) 2018/858 pentru vehicule cu motor
7. Reg. (EU) 2018/1139 pentru aviația civilă (e.g. drone)

Arii cu risc ridicat. Orice aspect din această categorie va fi evaluat înainte de a fi introdus pe piață și ulterior pe parcursul întregului ciclu de viață. Sunt enumerate opt arii de risc:

Infrastructură critică: sisteme de IA destinate a fi utilizate drept componente de siguranță în gestionarea și operarea traficului rutier și în aprovizionarea cu apă, gaz, încălzire și energie electrică

Educație și formare profesională: sisteme de IA destinate a fi utilizate în scopul stabilirii accesului sau al repartizării persoanelor fizice la instituțiile de învățământ și de formare profesională, sisteme de IA destinate a fi utilizate în scopul evaluării elevilor din instituțiile de învățământ și formare profesională și al evaluării participanților la testele necesare în mod normal pentru admiterea în instituțiile de învățământ

Ocuparea unui loc de muncă: gestionarea lucrătorilor și accesul la activități independente

Servicii publice: Accesul la servicii esențiale private și la serviciile și beneficiile publice, precum și posibilitatea de a beneficia de acestea

Aplicarea legii: care poate interfera cu drepturile fundamentale (e.g. analiza și evaluarea automatcă a dovezilor)

Controlul frontierelor: Gestionarea migrației, a cererilor de azil și a controlului la frontiere

Justiție: Administrarea justiției și procesele democratice

■ **Exemplu 1.16 — Risc ridicat - gestionarea infrastructurilor.** Aplicația IA care controlează sistemul de semafoare dintr-o intersecție. ■

■ **Exemplu 1.17 — Risc ridicat - ocuparea unui loc de muncă.** Utilizarea unui software de sortare a CV-urilor în procedurile de recrutare. ■

■ **Exemplu 1.18 — Risc ridicat - formare profesională.** Utilizarea unui software antiplagiat pentru verificarea originalității într-o lucrare de licență ar putea intra în această categorie. Ca urmare astfel de aplicații vor necesita un proces de certificare. ■

■ **Exemplu 1.19 — Risc ridicat - formare profesională.** Utilizarea unui software pentru evaluarea automată a testelor grilă utilizate la admiterea în învățământul superior. ■

■ **Exemplu 1.20 — Risc ridicat - aplicarea legii.** Aplicație IA care evaluează gradul de risc a unui condamnat pentru a recidiva, folosit ca suport decizional în vederea eliberării condiționate. ■

■ **Exemplu 1.21 — Risc ridicat - accesul la servicii critice.** Aplicație IA care prioritizează pacienții în sala de urgențe sau la transplant de organe. O astfel de aplicatie aduce valoare adăugată chiar daca performața sa nu este optimală. Criteriul de acceptare se rezumă la faptul că beneficiile cu introducerea IA sunt mai mare decât fără introducerea aplicației. ■

■ **Exemplu 1.22 — Risc ridicat - aplicarea legii.** Sistem de recunoaștere a emoțiilor folosit în aplicarea legii, cum ar ”detectorul de minciuni” pe baza d emiscarea ochilor, schimbări în voce, postura sau gesturi faciale. ■

■ **Exemplu 1.23 — Risc ridicat - ocuparea unui loc de muncă.** Sistem de recunoaștere a emoțiilor folosit pentru evaluarea candidaților în timpul interviurilor pentru angajare ■

■ **Exemplu 1.24 — Risc ridicat - ocuparea unui loc de muncă.** Sistem de sortare a candidților pentru interviu pe baza fotografiilor din CV ■

R Deși sistemele de recunoaștere a emoțiilor sunt în mod predefinit încadrate în categoria de risc limitat, deoarece aria de aplicabilitate este “ocuparea forței de muncă” sau “aplicarea legii“ exemplele 1.22, 1.23 sau 1.24 se incadrează în grupa aplicațiilor cu risc ridicat.

■ **Exemplu 1.25 — Risc ridicat - educație.**
Sistem IA pentru detectarea fraudei în timpul examenelor prin monitorizare video (e.g. Ghostwriter).
■

Cerințe pentru sistemele RBI
Gestiunea riscurilor: stabilirea unui sistem de gestiune a riscurilor pe întreg ciclul de viață al produselor
Date: utilizarea unor seturi de de date de antrenament și testare calitative
Documentație: tehnică plus capabilități de logare pentru facilitarea trasabilității și a auditului
Transparență: informarea utilizatorilor cu privire la capabilitățile și limitările sistemului de identificare
Supraveghere umană: e.g. identificările semnalizate de sistemul IA sunt verificate manual de cel puțin două persoane (principiul ”celor patru ochi“)
Robustețe, acuratețe, securitate: prin stabilirea de praguri de performanță pentru aceste criterii.

Cadrul de reglementare AIA fiind construit pe linia NLF se bazează pe standarde specifice pentru date biometrice:

- ISO/IEC 19794-5:2005 - specifică cerințele de tehnice (scenă, fotografie, digitizare, format) pentru imagini care conțin fețe utilizat eîn contextul verificarii umane sau recunoașterii automate
- ISO/IEC TR 22116 - analiza impactului factorilor demografici în performanța unui sistem de recunoaștere pe bază de date biometrice.

Obligațiile operatorilor de sisteme RBI
- Stabilirea unui sistem de gestiune a riscurilor

- Redactarea documentației tehnice
- Obligații de logare pentru a permite monitorizarea de către un agent uman a sistemului
- Evaluarea sistemului de către o parte terță
- Înregistrarea sistemului în baza de date europeană a aplicațiilor IA cu risc ridicat
- Obligativitatea monitorizării produsului după lansarea pe piața
- Semnarea declarației de conformitate
- Colaborarea cu autoritățile de supraveghere a pieței

Astfel, pe baza acestor riscuri, furnizorii de sisteme de IA cu risc ridicat vor fi supuși unor etape de evaluare și unor obligații stricte, înainte ca produsul să fie introdus pe piață (Figura 7.4).

R Statele membre s-au angajat să protejeze democrația și drepturile fundamentale ale cetățenilor. Ele au datoria să respecte aceste angajamente și în raport cu riscurile introduse de inteligența artificială.

Cerințe pentru sisnteme IA cu risc ridicat (72)

Gestiunea datelor	Asigurarea unui set de criterii de calitate pentru antrenarea, validarea și testarea modelelor care folosesc învățare automată
Documentație tehnică	Realizarea acesteia înainte de lansarea produsului pe piață și introducerea în documentației a dovezilor că sistemul este conform cu cerințele AIA
Înregistrarea operațiunilor	Dezvoltarea aplicației cu capabilități de logare automată a tuturor evenimentelor
Transparență și informarea utilizatorilor	Ieșirea aplicației IA trebuie să poată fi interpretată de agenul uman.
Supraveghere umană	Proiectarea aplicației astfel incât să poată fi supravegheță și controlată de agentul uman
Acuratețe, robustețe și securitate	Funcționarea în mod consitent la parametrii specificați de acuratețe, robustețe și securitate
Gestiunea riscurilor	Stabilirea și implementarea unui sistem de gestiune a riscurilor
Gestiunea calității	Asigurarea conformanței cu AIA și stabilirea unui sistem de monitorizare a aplicației după lansarea pe piață

1.3 Sisteme IA cu risc limitat și risc scăzut

Risc limitat. Sistemele de IA, cum ar fi cele de tip chatbot, fac obiectul unor obligații minime de transparență menite să le permită celor care interacționează cu un astfel de conținut să ia decizii în cunoștință de cauză, astfel utilizatorul să decidă ulterior dacă va continua sau nu să utilizeze aplicația.

■ **Exemplu 1.26** Risc limitat: recunoașterea emoțiilor Sistem IA într-un magazin care identifică reacțiile clienților la reclame pe baza expresiilor faciale ■

■ **Exemplu 1.27** Risc limitat: recunoașterea emoțiilor Sistem de detecție instalat într-un vehicul pentru a detecta starea de oboseală a șoferului și activa alerte specifice. ■

Aplicațiile destinate clasificarii biometrice se încadrează în grupa de risc limitat. Prin clasificarea biometrică se înțelege asignarea persoanelor unor categorii specifice în funcție de sex, vârstă, culoare ochilor, origine etnica, orientare sexuală pe baza datelor biometrice.

■ **Exemplu 1.28** Risc limitat: clasificare biometrică Aplicație IA care prezice dacă o persoană prezice homesexualitatea. ■

■ **Exemplu 1.29** Risc limitat: clasificare biometrică Un furnizor de jocuri video care utilizează clasificarea biometrică pentru a verifica vârsta utilizatorilor. ■

Risc minim. Marea majoritate a sistemelor de IA se încadrează în această categorie.

Reglementările pentru IA se vor sprijini sau vor completa reglementările existente în materie de siguranță. De exemplu, raportul privind implicațiile în materie de siguranță și răspundere ale inteligenței artificiale, ale internetului obiectelor și ale roboticii [24] concluzionează că normele pentru verificarea produselor asimilate mașinilor nu acoperă toate aspectele pe care IA le ridică. În propunerea de regulament al Parlamentului European și al Consiliului privind produsele asimilate mașinilor [27] sunt enumerate riscuri pe care produsele cu IA le conțin, altele decăt siguranța fizică.

R Conform directivei UE pentru produsele asimilate mașinilor, producătorul este responsabil pentru siguranța produsului pus pe piață, inclusiv a părților sale componente, e.g. software

Prin acest cadrul, CE urmărește armonizarea legislației cu privire la IA evitarea situațiilor in care fiecare stat membru va avea reglementari eteorgene. Valoarea adaugată în cadrul UE prin abordarea unitară la nivel statelor membre a reglementărilor pentru IA este detaliata în [35].

R O abordare comună a reglementărilor pentru IA are potențial de a genera până la 294.9 miliarde Euro suplimentare la GDP și 4.6 milioane de locuri de muncă până în 2030 [35].

Cadrul normativ pentru IA se află la intersecția dintre pe de o parte drepturile omului, eticii, cadre naționale și internaționale, iar pe de altă parte libertatea de inovare.

Documente relevante pentru reglementarea IA includ: [43], [18], [15], [17], [21], [20], [25], [26], [19], [3], [10]. [a] [b]

[a] https://epthinktank.eu/2020/11/23/what-is-the-european-parliaments-position-on-artificial-intelligence/
[b] Inception impact assessment

2. Politici emergente pentru reglementarea IA

Rezumat. Capitolul prezintă instituții create în diferite state pentru monitorizarea inteligenței artificiale sau pentru implementarea strategiilor naționale pe IA. Abordările naționale sunt diferite, în funcție de percepția riscurilor și a beneficiilor asociate cu IA, în funcție de domeniul de aplicabilitate.

2.1 Riscurile asociate inteligenței artificiale

Exista exemple de cazuri în care aplicațiile IA au afectat siguranța persoanelor sau drepturile fundamentale ale acestora. Aceste exemple ajută la: (i) ilustrarea motivației pentru care cadrele de reglementare sectoriale nu sunt suficiente; (ii) întelegerea manierei de aplicabilitate a cadrului de reglementare în situații viitoare.

Directive (EU) 2019/1024. - directiva datelor deschise și a reutilizării datelor în sectorul public
Regulation (EU) 2018/1807 flux de date non-personale în EU
European Strategy for Data COM (2020) 66

Identificare Biometrică În 2020 lanțul de magazine Mercadona din Spania a început utilizarea în cele 40 de magazine ale sale a unui sistem de recunoaștere facială pentru a identifica ”persoanele cu ordine de restricție și a celor cu cazier“. Investigația Agenției Spaniole pentru Protectia Datelor a condus la amendarea Mercadona cu 2,252,000 Euro (PS/00120/2021) deoarece identificarea facială s-a efectuat pentru toate persoanele aflate in magazine (clienți, angajați, copii). În plus, cerințele de transparență impuse de Articolele 12 și 13 din GDPR au fost de asemenea violate.

2.2 Instituții și politici pentru inteligența artificială

Țările urmăresc abordări naționale de guvernanță variate pentru a coordona punerea în aplicare a strategiilor și politicilor lor naționale de IA în toate guvernele, oferind supraveghere etică și de reglementare. Pentru a asigura coerența politicilor și implementarea eficientă a politicilor naționale de IA, guvernele utilizează patru modele [33]:

1. Atribuirea supravegherii dezvoltării și implementării unei strategii unui minister sau agenție deja existente;

- Biroul de politici științifice și tehnologice al Casei Albe[1] supervizează strategia națională a SUA pentru IA
- Ministerul Afacerilor Economice și al Comunicatiilor din Estonia a creat strategia nationala a Estoniei pentru IA
- Biroul primului ministru coordoneaza implementarea politicilor IA în Franța

2. Crearea unui nou organism guvernamental sau de coordonare pentru IA;
   - Politica IA în Marea Britanie este coordonată de Biroul guvernului britanic pentru IA
   - Casa Albă a SUA a înființat Biroul Național de Inițiativă IA, respectiv Sub-Comitetul de Învățare automata și Inteligență Artificială (MLAI)[2] [49]
   - Singapore a creat un Birou Național de IA pentru a coordona implementarea strategiei sale naționale de IA
3. Înființarea unor grupuri consultative de experți în IA
   - Consiliul austriac pentru robotică și IA
   - Consiliul consultativ canadian pentru IA
   - Consiliul consultativ al Spaniei pentru IA
   - Comitetul selectat al SUA pe IA în cadrul Consiliului Național de Știință și Tehnologie
4. Primirea de informații de la organele de supraveghere și cele consultative pentru IA și organismele de etică a datelor
   - Comitetul de etică al datelor din Germania
   - Grupul consultativ pentru etica datelor din Noua Zeelandă
   - Centrul pentru Etică a Datelor și Inovare din Marea Britanie (CDEI)
   - Consiliul consultativ pentru utilizarea etică a IA și a datelor din Singapore

Special Committee on Artificial Intelligence in a Digital Age

Instituții pentru dezvoltarea și implementarea strategiilor naționale

Printre ministerele sau agențiile existente însărcinate cu dezvoltarea sau implementarea unei strategii de IA, cele care tind să conducă cel mai adesea la crearea strategiilor de IA sunt i) ministerele tehnologiei informației și comunicațiilor; ii) ministerele economiei sau finanțelor; iii) ministerele educației, științei (și tehnologiei) și inovării.

SUA: Cele trei strategii pentru IA ale SUA, au fost redacte de MLAI ("Preparing for the Future of Artificial Intelligence" [37]), NITRD[3] ("National Artificial Intelligence R&D Strategic Plan"), respectiv de către Executive Office of the President ("Artificial Intelligence, Automation, and the Economy" [83])

Brazilia: Eforturile naționale de strategie IA au fost conduse de Ministerul Științei, Tehnologiei, Inovațiilor și Comunicării

Estonia: Ministerul Afacerilor Economice și Comunicațiilor a creat strategia națională de IA

India: Strategia IA este supravegheată de Ministerul Planificării, un minister axat pe încurajarea cooperării între statele indiene

Israel: Strategia IA a fost dezvoltată prin intermediul echipei sale guvernamentale, un efort interdepartamental cu membri din Biroul primului ministru, Ministerul Apărării, Autoritatea de Inovare a Israelului, Direcția cibernetică națională și Consiliul pentru învățământul superior

Polonia: Strategia IA a Poloniei a fost dezvoltată de Ministerul Afacerilor Digitale, Ministerul Dezvoltării și Tehnologiei, Ministerul Științei și Ministerul Fondurilor Publice

Rusia: Implementarea strategiei naționale IA este supravegheată de Ministerul Dezvoltării Economice.

[1] White House Office of Science and Technology Policy (OSTP)
[2] Machine Learning and Artificial Intelligence Sub-Committee (MLAI)
[3] Network and Information Technology R&D

R Implementarea strategiei naționale de IA variază de la o țară la alta. Unele țări au atribuit la mai mult de un minister responsabilitatea pentru coordonarea politicilor IA.

▪ Exemplu 2.1 — Responsabilitate interministerială pentru IA.

- Strategia IA a Germaniei a fost un proiect comun între ministerele federale ale educației și cercetării, afacerilor economice și energiei și muncii și afacerilor sociale
- Ministerul Industriei și Tehnologiei și Biroul de Transformare Digitală al Președinției lucrează împreună la dezvoltarea strategiei naționale de IA a Turciei.

▪

R Factorii cheie pentru implementarea eficientă a politicii IA sunt:

1. Sprijinul la nivel de lider
2. Asigurarea coordonării orizontale.

▪ Exemplu 2.2 — Inițiative naționale pentru strategia IA. Repatriot consideră în (87, page 2) patru obiecte strategice ale României privind tehnologia IA

- De a crește calitatea vieții oamenilor
- De a facilita dezvoltarea economică
- De a sprijini Europa să devină un lider în inovare digitală
- De a contribui la piața globală, cu soluții IA performante și sigure.

▪

Dezvoltarea politicilor naționale care se concentrează în mod special pe IA este un fenomen relativ nou. Există cinci recomandări pentru guverne [33]:

1. Investiții în cercetare și dezvoltare IA
2. Încurajarea unui ecosistem digital pentru IA
3. Formarea unui mediu politic favorabil IA
4. Dezvoltarea capacității umane și pregătirea pentru transformarea pieței muncii
5. Încurajarea cooperării internaționale pentru IA de încredere.

R Reglementările pentru IA ar trebui să acopere două categorii de riscuri: (1) riscurile pentru drepturile fundamentale, respectiv (2) riscurile dlegate de siguranța fizică și psihică a persoanelor.

R Ar fi utile politici pentru sprijinirea partajării datelor și asigurarea de protocoale și spații pentru testare și experimentare.

Portofoliu de colaborări IMM - multinaționale Firmele mici cu inovare pot beneficia de pe urma expertizei și aparatului de promovare și comercializare ale companiilor mari.

AIA fiind un regulament va avea efecte de aplicare directă pentru statele membre. Anumite aspecte sunt lăsate explicit la decizia fiecarui stat membru. Aceste aspecte ar putea fi subiectul unei consultări publice pe tema reglementării inteligenței artificiale.

Consultare publică pe reglementarea IA

- Cuantumul amenzilor pentru punerea pe piață sau utilizarea unei aplicații IA considerate interzise să fie 6% din cifra de afaceri (maxim 30 milioane Euro)?
- Identificarea biometrică în timp real să fie considerată o practică interzisă inclusiv în spațiile publice virtuale/online (i.e nu doar în spațiile publice fizice)?

R Tendința generală este ca politicile de acces și partajare a datelor să fie legate tot mai mult de politicile care privesc direct inteligența artificială.

■ **Exemplu 2.3 — Încurajarea unui ecosistem digital pentru IA.** Proiectul OECD privind îmbunătățirea accesului și schimbului de date (EASD)

OCDE a analizat modul în care îmbunătățirea accesului la date poate maximiza valoarea socială și economică a datelor. Raportul din noiembrie 2019 „Enhancing Access to and Sharing of Data: Reconciling Risks and Benefits for Data Re-use across Societies" [**FG-AI4HP**] identifică cele mai bune practici pentru a echilibra diferitele interese pentru a profita de avantajele accesului și partajării datelor, gestionând în același timp și reducând riscul la un nivel social acceptabil. Instrumentele juridice includ:

- OECD (2006, actualizată în 2021) Recomandarea Consiliului privind accesul la datele de cercetare din finanțarea publică
- OECD (2008) Recomandarea Consiliului pentru acces sporit și o utilizare mai eficientă a informațiilor din sectorul public
- OECD (2014) Recomandarea Consiliului privind strategiile guvernamentale digitale
- OECD (2016) Recomandarea Consiliului privind guvernarea datelor privind sănătatea.

■

R Adoptarea necesită acces la tehnologiile și la capacitatea de calcul IA și investirea în tehnologii lingvistice care acoperă diverse discipline de IA (precum procesarea limbajului natural (NLP), extragerea informațiilor, recunoașterea vorbirii, lingvistică de calcul și traducere automată) [33]

■ **Exemplu 2.4 — Prioritizări ale acestor tehnologii în strategiile de IA.** sisteme interactive de dialog și asistenți virtuali personali pentru servicii publice personalizate, servicii de traducere automată care ar putea atenua barierele lingvistice din comerțul electronic internațional, în special pentru IMM-uri, consolidarea de seturi de date pentru a crea resurse pentru antrenarea sistemelor IA bazate pe tehnologii lingvistice. ■

Cadre de reglementare pentru IA de încredere

Diverse țări explorează abordări pentru a asigura IA de încredere și pentru a atenua riscurile asociate cu dezvoltarea și implementarea sistemelor de IA. Acțiunile de reglementare emergente pentru încrederea în IA includ:

- explorarea aplicației
- necesitatea de a adapta legislația actuală pentru IA
- furnizarea de îndrumări legale
- luarea în considerare a abordărilor stricte ale legii
- introducerea unor interdicții specifice aplicației
- promovarea mediilor controlate pentru experimentarea reglementării
- sprijinirea eforturilor internaționale de standardizare și a eforturilor de drept internațional.

Clasificarea sistemelor IA conform OECD (33) în funcție de impactul asupra politicilor publice cuprinde:

1. Contextul în care funcționează sistemul (de exemplu, sectorul aplicației, amploarea implementării)
2. Datele și intrările utilizate de sistem (de exemplu, calitatea, confidențialitatea datelor)
3. Modelul IA care stă la baza sistemului (de exemplu, modele generative sau simbolice)
4. Sarcina și rezultatul pe care le produce sistemul (de exemplu, nivelul de autonomie, natura rezultatului).

Liniile directoare pentru IA de încredere oferă standarde pentru utilizarea etică a IA și guvernarea acesteia. În funcție de caz, acestea se adresează factorilor de decizie politică, întreprinderilor, instituțiilor de cercetare și altor actori IA.

Recomandări și ghiduri voluntare

- Cadrul etic pentru IA al Australiei
- Instrumentul de autoevaluare online al Belgiei pentru a încuraja IA de încredere, adaptat în mod special sectorului public
- Cadrul de etică al Columbiei pentru inteligența artificială
- Carta Egiptului privind IA responsabilă include orientări de evaluare, orientări tehnice și bune practici
- Liniile directoare etice ale Ungariei
- Liniile directoare japoneze privind cercetarea și dezvoltarea si liniile directoare privind utilizarea IA
- Cadrul de explicabilitate a IA al Scoției

R La nivelul UE, grupul independent de experți la nivel înalt (AI HLEG) al Comisiei Europene a introdus Orientări etice privind IA în decembrie 2018. În iulie 2020, AI HLEG a prezentat o listă de evaluare pentru inteligența artificială de încredere [22].

R Asumpția de simetrie comportamentală [5] - atât actorii din business (i.e. realizatorii și implementatorii de coduri etice), cât și actorii din stat (i.e ca reglmentare și impememntare de legi) au același risc de a avea un comprtament moral sau nu. (Persoanele implicate in reglementări la nivel de stat sunt la fel ca în orice instituție privată.)

Spațiile de testare în materie de reglementare a IA vizează în paralel: abordări de co-reglementare care urmăresc permiterea experimentării pentru a înțelege mai bine efectele sistemelor de IA și a oferi medii controlate pentru a facilita extinderea noilor modele de afaceri. Aceste abordări de reglementare ajută la crearea unui mediu care sprijină tranziția de la cercetare la implementarea unor sisteme de încredere în IA [32], [81].

Conceptul de spațiu de testare (i.e. sandbox) a fost introdus în Statele Unite, urmate de Autoritatea de Conduită Financiară din Regatul Unit. Obiectivul acestor spații de testare a fost de a testa noi produse și servicii Fintech înainte de a intra oficial pe piață.

În ianuarie 2021, nu existau instrumente de guvernanță obligatorii pentru a reglementa în mod specific sistemele de IA. Cu toate acestea, mai multe guverne și organisme interguvernamentale au adoptat sau iau în considerare legislația obligatorie pentru domenii specifice ale tehnologiilor IA.

Abordări de tip hard law

- Direcția rutieră daneză a emis un ghid obligatoriu pentru mașinile fără șofer
- Iunie 2017, Germania le-a permis șoferilor să transfere controlul vehiculelor către sisteme de conducere superior sau complet automatizate pentru utilizare pe drumurile publice
- În Statele Unite, Administrația Federală a Aviației a implementat noi reglementări și programe pilot pentru accelerarea integrării sistemelor de aeronave fără pilot în sistemul național de spațiu aerian
- În 2020, Administrația SUA pentru Alimente și Medicamente (FDA) a luat în considerare reglementarea anumitor sisteme de diagnostic medical alimentat de IA
- Februarie 2020, New York a introdus un regulament privind „vânzarea instrumentelor automate de decizie a ocupării forței de muncă”

Interdicții directe/efective pot fi puse în aplicare de către guverne pentru a menține mecanismele pieței existente sau pentru a proteja cetățenii de consecințele negative ale tehnologiilor IA:

- Belgia a adoptat rezoluții pentru a interzice utilizarea armelor autonome letale de către forțele armate locale
- În ultimii ani, tehnologia de supraveghere biometrică sau de recunoaștere facială a apărut ca o problemă importantă a dezbaterii publice. Riscurile de prejudecăți algoritmice necorespunzătoare și preocupările privind confidențialitatea datelor au dus la diverse apeluri și acțiuni de interzicere a utilizării tehnologiei de recunoaștere facială. În Statele Unite, atât guvernele federale, cât și guvernele de stat au indicat dorința de a promulga reglementări privind utilizarea tehnologiei de recunoaștere facială de către agențiile guvernamentale sau forțele de ordine, ca o politică inovatoare și un instrument de reglementare pentru îmbunătățirea încrederii de reglementare asupra diferitelor daune și beneficii generate de emergențe, modele de afaceri și tehnologii.

Proprietăți pentru XAI În august 2020, NIST a publicat o lucrare care cuprinde patru proprietăți fundamentale pentru sistemele XAI [80]:
Explicație: Furnizarea dovezilor însoțitoare sau motive pentru toate rezultatele sistemului IA
Semnificativ: Furnizarea de explicații care să fie semnificative sau ușor de înțeles pentru utilizatorii
Exactitatea explicației: Explicația trebuie să reflecte corect procesul de generare a rezultatului
Limite de cunoaștere: Sistemul ar trebui să funcționeze numai în condițiile pentru care a fost proiectat sau atunci când sistemul atinge suficientă încredere în decizia/acțiunile sale.

R Danemarca deține secretariatul pentru identificarea nevoilor de standardizare în IA. Standardizarea în IA este văzută ca bază pentru viitoarele reglementări în domeniu.

Posibil impact al IA asupra drepturilor fundamentale:

1. Discriminare și bias
2. Manipularea opiniilor - afectarea autonomiei și a "human agency" prin companii de dezinformare, știri fabricate, motoare de recomandare
3. Libertatea de expresie și informare, dreptul la alegeri libere: probleme de cenzură sau filtrare informații, manipularea pe social media
4. Protectia datelor/privacy: recunoaștere facială, identificare biometrică
5. Guvernanta și dreptul la un proces echitabil - poliție predictivă (functionează deja în jumatate din țările UE), modelarea riscului, evaluarea comportamentului social
6. Protecția consumatorului, libertatea de asociere.

Danemarca este prima țară care a introdus cerințe obligatorii pentru companii conform cărora acestea să se conformeze principiilor etice în IA (datând din 2020 cu intrare în vigoare din 2021). Reglementările introduse urmează strategia Danemarcii pentru IA și urmăresc menținerea încrederii în țară și în eticheta "Made in Denmark", inclusiv în domeniul IA. Modificările legislative s-au produs printr-un amendament la Danish Financial Statements Act. Agenții economici trebuie să furnizeze informații despre algoritmii utilizați și sa demonstreze ca aceștia respectă cerinte de transparență. În raportul financiar anual se adaugă și o secțiune dedicată eticii datelor. Dezvoltatorii de aplicații IA care urmează șase principii pentru etica datelor pot să utilizeze pe pagina web o etichetă specifică pentru Data Ethics. Dintre aceste șase principii enumerăm:

- Auto-determinare: asigurarea că cetațenii pot lua decizii informate și independent
- Egalitate de sanse: asigurarea ca drepturile fundamentale nu sunt încălcate și manifestarea respectului pentru diversitate
- Deschidere și transparență: include responsabilitate în funrizare de explicații.

Legislatia pe etica datelor și securitate include inițiative precum Danish cyber and information security

strategy
În colaborare cu Danish Technical University s-au enunțat bune practici pentru aplicații IA sigure (Safe AI):
Securitate: Robustețe la atacuri
Open source: Metodele, codul sursă și rezultatele testelor sunt publice
Auto-cunoastere: Refuzul sistemului de a acționa în caz de incertitudine
Confidențialitate: Dezvoltare pe baza principiilor "privacy by design"
Valori calibrate: Proiectare și testare împotriva stereotipurilor sau biasului și înțelegerea emoțiilor
Responsabilitate: Comunicare transparentă în cazul activării "dreptului la explicație"
Înțelegerea relațiilor sociale: În completarea cunoștințelor și competențelor utilizatorilor
Înțelegerea puterii: Înțelegerea datelor, contextului și a consecințelor acțiunilor și deciziilor.

R În Danemarca, Consiliul pentru Etica Datelor a lansat în 2019 un set de unelte pentru etica datelor cu scopul de a ajuta companiile să implementeze principii de etică a datelor în procesele lor de business. De asemenea, s-a lansat eticheta D-seal pentru semnalizarea aplicațiilor care urmează princiipii de etică a datelor.

R Din 2017, codul rutier danez a fost extins pentru a permite testarea vehiculelor autonome.

R Articolul 25 din GDPR este intitulat “Data protection by design and by default”.

Australia: Cadru voluntar, 8 principii pentru IA etică, prioritate națională: dezvoltarea de standarde pentru IA
Canada: Una din primele țări care a introdus reglementări (e.g. Directive on Automated Decision-Making, 2018, 6 grupe de risc)
Germania: Data Ethics Commission, 5 grupe de risc
Japonia: Contract Guidelines on Utilisation of AI and Data - centrată pe modelarea contractelor pentru utilizarea datelor și a produselor IA
Singapore: Politici voluntare, implementarea unui Self-Assessment Guide for Organisations, inițierea autorităților în IA, determinarea nivelului de implicare a agentului uman în decizii propuse de IA, asigurarea interacțiunii actorilor, promovarea opțiunilor de opt-out pentru utilizatori
UK: Ghid pentru utilizarea IA în administrația publică - include recomandari legate de tehnicile de învatare automată pentru anumite procese administrative), UK’s Information Commissioner’s Office Guidance on AI Auditing - cadru voluntar pentru evaluarea riscurilor legate de protecția datelor personale, măsuri pentru securitate
US: Guidance for Regulation of Artificial Intelligence Applications - standarde voluntare de evaluare a conformitătii; New Jersey proposed Algorithmic Accountability Act.

R Aplicarea IA în toate sectoarele va atrage o schimbare culturală atât la nivelul dezvoltatorilor, cât și la nivelul utilizatorilor.

2.3 Coduri și reglementări în industrie

Acțiuni propuse de AlgorithmWatch:
1. Orice decizie în administrație care a fost luată pe baza unor informări/ieșiri de la IA cu privire la o persoană, aceasta trebuie informată
2. Procedura pentru contestare: orice individ poată să obțină informațiile relevante pe baza cărora s-a

luat decizia. Persoanele în cauză au afectate dreptul de a inspecta sistemele de decizie automate, documentația acestora și protocoalele.
3. Introducerea de obligații legale pentru publicare datelor (e.g. prin API publice)

Bias de confirmare reprezintă atrofierea abilităților persoanelor care interacționează și își bazează deciziile pe aplicații IA (e.g. de-skilling, hyper-nudging). S-a constatat ca deciziile recomandate de aplicațiile IA sunt acceptate în aproape toate cazurile de către agentul uman. Acest tip de bias cognitiv a fost constatat la diferite categorii profesionale: piloți de avion, judecători, medici.

■ **Exemplu 2.5 — Bias de confirmare - educație.** Un soft antiplagiat bazat pe procesare de limbaj natural semnalizează un procentaj de 30% similaritate între lucrarea de licență a unui student și surse de pe Internet. Comisia decide respingerea candidatului fără a verifica corectitudinea recomandării aplicației IA. ■

■ **Exemplu 2.6 — Evitarea biasului de confirmare în domeniul medical.** Două posibile soluții sunt:

1. Utilizatorii unui sistem de diagnostic medical să fie informați că sistemul este setat să dea un diagnostic cu totul aleator într-un numar de 10% din cazuri. Astfel se elimină tendința de conformare unitară cu diagnosticul furnizat de aplicația IA;
2. Stabilirea unui protocol de diagnostic om-mașină în care diagonosticul este dat întâi de agentul uman, apoi de aplicația IA.

■

■ **Exemplu 2.7 — Bias de gen.** Algoritm de recrutare care analizează CV-urile candidaților și acordă în mod consecvent scor mult mai mic femeilor. ■

■ **Exemplu 2.8 — Bias de rasă.** Algoritmii de recunoaștere a fețelor din imagini (Microsoft, Amazon, etc), care au o performanță extrem de redusă pentru anumite rase, în special cea afro-americană. ■

■ **Exemplu 2.9 — Bias de rasă.** Sistemul COMPAS folosit în sistemul de justiție din SUA pentru a prezice riscul de recidivism,; Sistmeul COMPAS producea pentru persoanele de culoare de două ori mai multe etichetări "high-risk" în situații în care nu au existat ulterior recidive. ■

Experimentul "Mașina morală" Un experiment social, pentru a ajuta la rezolvarea situațiilor limita ce pot apărea în cazul vehiculelor autonome într-o maniera compatibilă cu judecata umană.

Microsoft susține crearea unui cadru strategic de reglementare pentru dezvoltarea responsabilă a IA, pe linia documentelor UE. Spre exemplu, inițiativa Tech Fit 4 Europe prezinta politici emergente în UE legate de IA. Principiile Micorsoft privind inteligența artificială (transparența, corectitudinea, responsabilitatea, confidențialitatea și securitatea) stabilesc praguri de pretecție pentru aplicațiile IA. Aceste principii sunt operaționalizate de către trei grupuri: (1) Comitetul Aether care oferă exeprtiză legată de tendințele IA; (2) Biroul pentru Inteligență Artificială responsabilă care enunță politici, respectiv (3) Grupul pentru Strategia IA care monitorizează implementarea IA responsabilă.

Cine este raspunzător?

- White Paper on AI and the accompanying Report on Safety and Liability,
- Product liability Directive 85/374/EEC[a]
- Regulation on liability for the operation of Artificial Intelligence-systems[b]

[a] https://www.europarl.europa.eu/doceo/document/TA-9-2020-0276_EN.html
[b] European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civilliability regime for artificial intelligence (2020/2014(INL

Tabela 2.1: Argumente și contra argumente din spațiul public pentru relementarea IA

Argumente pro reglementari IA	**Argumente contra reglementărilor pe IA**
Limitează/îngrădesc/evită posibilele abuzuri de către firme și instituții	Tehnologiile emergente/disruptive sunt greu de reglementat
Limitează îngrijorările legate de aspecte etice, drepturi fundamentale, protectia datelor	Amplifică birocrația
	Pot bloca inovația
Stabilesc standarde comune pentru industrie, îmbunătățind predictibilitatea	Pot afecta competitivitatea prin inegalitățile între țări cu reglementări diferite
Protejează utilizatorii aplicațiilor IA	Reglementările ar trebuie să fie ultimul remediu, doar în caz de pericol real
	Limitare la principii generice datorită scepticismului legat de orice reglementare

R Liabilitatea dezvoltatorului de aplicații IA trebuie complementată de liabilitatea operatorului aplicației.

R Regulile de liabilitate pentru IA nu limitează inovația, ci reprezintă instrumente pentru controlul riscurilor.

Cine e responsabil? Conform raportului COMEST al NU, deoarece roboții sunt clasificați ca produse tehnologice, intră sub incidența doctrinei de "neglijență" prin care producătorul sau vânzătorul sunt responsabili.

2.4 Percepția publică a reglementării IA

La nivel UE, marea majoritatea respondenților (90%) susțin introducerea reglementărilor pentru IA și roboți, precum și existenta unor reglementări la nivel supra-statal (96%). Sustinerea reglementărilor la nivel public depinde de domenii (e.g. protecția datelor, valori și principii, încredere, competitivitate, siguranță fizică, proprietate intelectuală), dar și de ariile de aplicabilitate (e.g. vehicule autonome, roboți medicali, drone, augmentare personală) [68].

Rezultate consultare publică pentru reglementări la nivel UE:

- 84% nu sunt ingrijorați de costurile impuse de reglementări, 7% menționează costurile de certificare, iar 7% sunt ingrijorați de birocrația rezultată în urma reglementărilor
- Pentru reglementarea IA este preferată abordarea pe bază de White Papers
- Sunt subliniate: (i) importanța dezvoltării de competențe în domeniu, (ii) facilitarea accesului la date, (iii) importanța susținerii parteneriatelor între IMM-uri, companii mari și mediul academic.

R De ce reglementarea IA acum, având în vedere că IA nu este ceva nou? Un exemplu ar fi legat de constatarea că folosim IA în domeniul medical fără verificare fără transpareță, deci fără a ne asigura că sănătatea pacientului nu este afectată.

R Miza AIA este de a evita ca fiecare stat membru să vină cu propriile reglementări.

R Existența unui cadrul de reglementare la nivelul UE poate proteja firmele de anumite decizii politice.

R Abordarea centrată pe om a UE reprintă un contract "social" între IA și umanitate în care agentul uman este în centru și la control (i.e., abordarea "human-in-command").

IRCAI - International Research Centre on Artificial Intelligence under the auspices of UNESCO Centrul UNESCO pentru IA

Alte documente relevante [49], [22], [9], [110], [78]

3. IA etică și de încredere ca bază a reglementări

Rezumat. UE dorește să fie liderul în dezvoltarea de IA responsabilă. Cadrul de reglementare propus [28] pune aspectele etice precum transparența algoritmilor, reducerea discriminării, supravegherea umană pe prima poziție. Astfel, principiile etice pentru inteligența artificială reprezintă fundamentul în jurul căruia se vor formaliza reglementările pentru IA. După prezentarea acestor principii, capitolul enumeră un set de ghiduri care se referă IA etică, IA responsabilă sau IA de încredere. Aceste ghiduri pot fi puncte de pornire pentru operaționalizarea inteligenței artificiale etice și în implementarea procedurilor pentru auditarea și certificarea aplicațiilor IA etice. De asemenea, sunt prezentate pe scurt politicile pentru IA etică a Google, Axon, DeepMing, Microsoft, sau Amazon, precum și unelte de IA dezvoltate recent pentru a sprijini IA etică, inclusiv pe linia Explainable AI.

3.1 Necesitate

Pe 21 aprilie 2021, Comisia Europeană a adoptat prima propunere cuprinzătoare de regulament privind inteligența artificială (IA) de natură să transforme Europa într-un centru global pentru o IA de încredere și să conducă eforturile globale de a stabili standarde în acest domeniu.

Propunerea prezintă norme armonizate proporționale și flexibile, bazate pe o abordare bazată pe riscuri, care abordează riscurile la adresa siguranței și a drepturilor fundamentale pe care le prezintă utilizările specifice ale tehnologiei IA.

Margrethe Vestager, Executive Vice-President for a Europe fit for the Digital Age, a declarat:

> *"În ceea ce privește inteligența artificială, încrederea este o necesitate ("must"), nu o opțiune dorită ("nice to have"). Cu aceste norme de referință, UE este vârful de lance al dezvoltării de noi norme globale, pentru a se asigura că IA poate fi de încredere. Prin stabilirea standardelor, putem deschide calea către tehnologia etică la nivel mondial și ne putem asigura că UE rămâne competitivă pe parcurs. Favorabile viitorului și favorabile inovării, normele noastre vor interveni acolo unde este strict necesar: atunci când siguranța și drepturile fundamentale ale cetățenilor UE sunt în joc."* Europe fit for the Digital Age, PR, 21 April 2021

Această propunere respectă angajamentul politic al președintelui von der Leyen, care a anunțat în orientările sale politice pentru Comisia Europeană 2019-2024 "O Uniune care depune eforturi

pentru mai mult" [21], prin care Comisia va prezenta legislație pentru o abordare europeană coordonată cu privire la implicațiile umane și etice ale IA. În urma acestui anunț, la 19 februarie 2020, Comisia a publicat „Cartea albă privind IA - O abordare europeană a excelenței și a încrederii" [26]. Cartea albă stabilește opțiunile de politică privind modul de realizare a dublului obiectiv de promovare a adoptării IA și de abordare a riscurilor asociate anumitor utilizări ale unei astfel de tehnologii.

Cartea albă privind IA (26). Comisia se angajează să permită progresul științific, să mențină poziția de lider tehnologic a UE și să se asigure că noile tehnologii sunt în serviciul tuturor europenilor, îmbunătățindu-le viața, respectând în același timp drepturile acestora Creșterea economică durabilă actuală și viitoare a Europei și bunăstarea societății se bazează din ce în ce mai mult pe valoarea creată de date. IA este una dintre cele mai importante aplicații ale economiei datelor. În prezent, majoritatea datelor sunt legate de consumatori și sunt stocate și prelucrate pe infrastructura centralizată bazată pe cloud. Strategia europeană pentru date, care însoțește prezenta Carte albă, urmărește să permită Europei să devină cea mai atractivă, sigură și dinamică economie agilă din lume în ceea ce privește datele, capacitând Europa cu date pentru a îmbunătăți deciziile și pentru a îmbunătăți viața tuturor cetățenilor săi.

Sintetic spus, IA este o colecție de tehnologii care combină datele, algoritmii și puterea de calcul. Progresele în materie de calcul și disponibilitatea tot mai mare a datelor sunt, prin urmare, factori-cheie ai creșterii actuale a IA. Europa își poate combina atuurile tehnologice și industriale cu o infrastructură digitală de înaltă calitate și cu un cadru de reglementare bazat pe valorile sale fundamentale pentru a deveni un lider mondial în domeniul inovării în economia datelor și în aplicațiile sale, astfel cum se prevede în Strategia europeană privind datele. Pe această bază, se poate dezvolta un ecosistem de IA care aduce beneficiile tehnologiei întregii societăți și a economiei europene:

- pentru ca *cetățenii* să poată avea noi beneficii, de exemplu îmbunătățirea asistenței medicale, mai puține defecțiuni ale mașinilor de uz casnic, sisteme de transport mai sigure și mai curate, servicii publice mai bune;
- pentru *dezvoltarea întreprinderilor/afacerilor*, de exemplu o nouă generație de produse și servicii în domenii în care Europa este deosebit de puternică (mașini, transport, securitate cibernetică, agricultură, economie verde și circulară, asistență medicală și acele sectoare cu valoare adăugată ridicată, cum ar fi moda și turismul);
- pentru *servicii de interes public*, de exemplu prin reducerea costurilor de furnizare a serviciilor (transport, educație, energie și gestionarea deșeurilor), prin îmbunătățirea sustenabilității produselor și prin punerea la dispoziția autorităților care răspund de aplicarea legii, instrumente adecvate pentru a asigura securitatea cetățenilor, cu garanții adecvate pentru a le respecta drepturile și libertățile.

Prezenta propunere vizează punerea în aplicare a celui de-al doilea obiectiv pentru dezvoltarea unui *ecosistem de încredere prin propunerea unui cadru juridic pentru o IA de încredere*. Propunerea se bazează pe valorile UE și pe drepturile fundamentale și urmărește să ofere cetățenilor și altor utilizatori încrederea de a adopta soluții bazate pe IA, încurajând în același timp întreprinderile să le dezvolte. IA ar trebui să fie un instrument pentru oameni și să fie o forță a binelui în societate, cu scopul final de a crește bunăstarea umană.

Normele privind IA disponibile pe piața Uniunii sau care afectează în alt mod persoanele din Uniune ar trebui, prin urmare, să fie *centrate pe om*, astfel încât oamenii să poată avea încredere că tehnologia este utilizată într-un mod sigur și conform cu legea, inclusiv cu respectarea drepturilor fundamentale.

> *"Fundamentul comun care unește drepturile fundamentale (demnitate, libertăți, egalitate și solidaritate, drepturile cetățenilor și justiție) poate fi înțeles ca având rădăcinile în respectul pentru demnitatea umană - reflectând astfel ceea ce descriem ca o abordare*

centrată pe om în care ființa umană se bucură de un statut moral unic și inalienabil de supremație în domeniile civil, politic, economic și social."

În urma publicării Cărții albe, Comisia a lansat o amplă consultare, întâmpinată cu un interes deosebit de un număr mare de părți implicate, care au sprijinit în mare măsură intervenția în materie de reglementare pentru a aborda provocările și preocupările ridicate de utilizarea tot mai frecventă a IA.

Merită subliniat faptul că accentul puternic pus pe etică în strategia Uniunii Europene în domeniul IA ar trebui privit în contextul unei strategii globale care vizează protejarea cetățenilor și a societății civile împotriva abuzurilor tehnologiei digitale, dar și ca parte a unei strategii orientate spre competitivitate, care vizează ridicarea standardelor de acces la piața unică a Europei. În acest context, unul dintre pașii specifici din strategia Uniunii Europene a fost crearea unui grup independent de experți la nivel înalt privind IA (AI HLEG), însoțit de lansarea unei alianțe pentru IA, care a atras rapid câteva sute de participanți. AI HLEG, un grup tip multistakeholder care include cincizeci și doi de experți, a fost însărcinat cu definirea orientărilor în materie de etică, precum și cu formularea "recomandărilor de politică și de investiții". Cu avizul AI HLEG, Comisia Europeană a prezentat orientări etice pentru o IA de încredere, care deschid acum calea către un cadru de politică cuprinzător, bazat pe riscuri [86].

Sumarizând, putem observa că obiectivul general al UE este de a crea o piață unică pentru IA în Europa, prin crearea încrederii necesare pentru ca utilizatorii să beneficieze de aplicații IA sigure și de încredere, oferind în același timp întreprinderilor și oricăror alți furnizori certitudinea juridică pentru a le inova și a le dezvolta. În mod evident, data de 21 aprilie, reprezintă un ”momentum“ în evoluția domeniului reglementării Inteligenței Artificiale în Uniunea Europeană.

Două concepte sunt esențiale în abordarea Uniunii Europene: *Excelență* și *Încredere*.

Strategia europeană în domeniul IA și planul coordonat clarifică faptul că încrederea este o condiție prealabilă pentru a asigura o abordare centrată pe om a IA: IA nu este un scop în sine, ci un instrument care trebuie să servească oamenii cu scopul final de a crește bunăstarea umană. Pentru a realiza acest lucru, ar trebui să se asigure încrederea în IA. Valorile pe care se bazează societățile noastre trebuie să fie pe deplin integrate în modul în care se dezvoltă și utilizează IA.

Trebuie subliniat și faptul că promovarea dezvoltării unei IA centrate pe om, durabilă, sigură, favorabilă incluziunii și de încredere este o provocare globală, iar dimensiunea internațională a IA este mai esențială ca niciodată. Implicațiile noilor tehnologii digitale, cum ar fi IA, depășesc frontierele și trebuie abordate la nivel mondial.

Organisme internaționale precum Organizația Națiunilor Unite pentru Educație, Știință și Cultură (UNESCO), Organizația pentru Cooperare și Dezvoltare Economică (OCDE), Consiliul Europei, G7 și G20 lucrează la aspecte conexe legate de IA.

Organizații precum Organizația Internațională de Standardizare (ISO) și Institutul inginerilor electrici și electronici (IEEE) sunt implicate în activități multiple de standardizare în domeniu. Comisia Europeană și Agenția pentru Drepturi Fundamentale, OCDE, Consiliul Europei, UNESCO și guvernul francez, în dublul său rol de viitoare președinție a Consiliului UE și de următorul președinte al Parteneriatului global pentru IA, colaborează în domeniul IA etice și de încredere; diferite organizații cooperează deja și discută cele mai importante aspecte pe care trebuie să le abordeze în comun, explorând sinergii suplimentare între diferitele direcții de lucru și posibilitățile de cooperare multilaterală consolidată.

Se poate remarca din toate aspectele prezentate mai sus, că încrederea este atât de importantă pentru domeniul IA, iar eforturile făcute sau planificate a se desfășura în perioada următoare se justifică prin impactul pe care IA îl poate avea în toate domeniile vieții economice, sociale, culturale și la nivel individual, cu condiția de a nu pune în pericol valorile și libertățile fundamental umane și, chiar existența umană.

Încrederea este condiție *sine-qua-non* pentru utilizarea aplicațiilor IA și, pe cale de consecință,

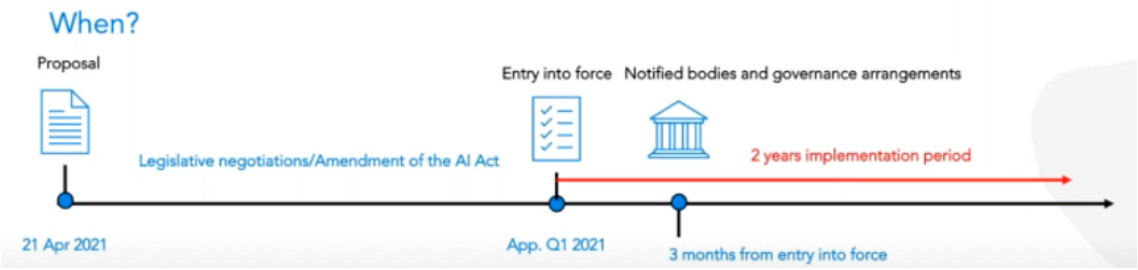

Figura 3.1: Pașii următori pentru reglementarea inteligenței artificiale

posibilitatea de a valorifica beneficiile acestora în folosul umanității. Ce urmează, la nivelul UE, se poate vedea în Figura 3.1.

Spre finalul acestui subcapitol, supunem atenției teza din Cartea albă, care merită citită și înțeleasă în esența ei, pentru a vedea spiritul și litera documentelor programatice ale UE, acțiunile propuse de UE și modul în care statele membre ar trebui să se raporteze și sa acționeze:

> *Este necesară o abordare europeană comună a domeniului IA pentru a atinge o masă critică suficientă și pentru a evita fragmentarea pieței unice. Introducerea inițiativelor naționale riscă să pună în pericol certitudinea (securitatea) juridică, să slăbească încrederea cetățenilor și să îngreuneze apariția unei industrii europene dinamice.*

R Analizele evaluării valorii adăugate europene (EAVA- European added value assessment) sugerează că un cadru comun al UE privind etica are potențialul de a aduce Uniunii Europene 294,9 miliarde EUR în PIB suplimentar și 4,6 milioane EUR în locuri de muncă suplimentare până în 2030; cred că este un argument important, care să conducă la acțiune sinergică din partea tuturor celor interesați/afectați [35].

Alte documente relevante: [57], [60].

3.2 Ghiduri pentru inteligența artificială etică

Inteligența artificială are potențialul de a sprijini realizarea unora dintre cele mai complexe probleme ale societății umane. În același timp, IA are potențialul de a perturba societățile prin impactul său asupra structurilor economice și sociale existente. Riscurile implicate în implementarea acestei tehnologii puternice includ o reducere a controlului asupra sistemelor digitale, introducerea de prejudecăți bazate pe gen sau rasă, precum și o creștere radicală a inegalității sociale, sau, după părerea unor experți, poate duce chiar la sfârșitul speciei umane. Riscurile menționate pot determina ca această dezvoltare tehnologică cheie cu un potențial impact pozitiv major pentru omenire, să fie anulată de frică și neîncredere [60].

Pentru a exploata oportunitățile și a preveni amenințările, este important să se sporească încrederea în IA și să se monitorizeze dezvoltarea acesteia. Pentru acest lucru sunt necesare ghiduri etice. În acest scop, au fost publicate coduri și principii etice [8] de guverne (de exemplu, Franța), de sectorul privat (de exemplu, Google, IBM) și de institutele de cercetare (de exemplu, Future of Life Institute).

Cu toate că este un acord cvasi-unanim că IA ar trebui să fie etică, există dezbateri cu privire la ceea înseamnă "IA etică" și la cerințele etice și standardele tehnice necesare pentru a o atinge.

Ghidul de etică pentru IA de încredere a fost prezentat de AI-HLEG pe 8 aprilie 2019 [34] și se bazează pe un proiect publicat în decembrie 2018 la care au fost formulate peste 500 de observații

în urma unei consultări publice. În conformitate cu orientările AI-HLEG, IA de încredere ar trebui să fie:

1. legală - cu respectarea tuturor legilor și reglementărilor aplicabile;
2. etică - respectarea principiilor și valorilor etice;
3. robustă - atât din punct de vedere tehnic, cât și ținând cont de mediul social în care se desfășoară.

Fiecare componentă în sine este necesară, dar nu suficientă pentru realizarea unei IA de încredere. În mod ideal, toate cele trei componente funcționează în armonie și se suprapun în funcționarea lor. Dacă, în practică, apar tensiuni între aceste componente, societatea ar trebui să depună eforturi pentru a le alinia[1].

Folosind o abordare bazată pe drepturile fundamentale, capitolul I (Foundations of Trustworthy AI) identifică principiile etice și valorile corelate ale acestora care trebuie respectate:

- Dezvoltarea, implementarea și utilizarea sistemelor IA într-un mod care să adere la principiile etice: respectarea autonomiei umane, prevenirea efectelor periculoase, echitatea și explicabilitatea; este important să fie recunoascute și abordate potențialele tensiuni dintre aceste principii;
- Să se acorde.o atenție deosebită situațiilor care implică categorii vulnerabile specifice, cum ar fi copiii, persoanele cu dizabilități și alte persoane care au fost dezavantajate în trecut sau sunt expuse riscului de excluziune, precum și situațiilor care se caracterizează prin asimetrii în capacitatea de informare, cum ar fi între angajatori și angajați, sau între întreprinderi și consumatori;
- Recunoașterea faptului că, deși aduc beneficii substanțiale persoanelor și societății, sistemele de IA prezintă, de asemenea, anumite riscuri și pot avea un impact negativ, inclusiv efecte care pot fi dificil de anticipat, de identificat sau de măsurat (de exemplu, asupra democrației, statului de drept și justiției distributive sau asupra minții umane în sine); de aici, necesitatea adoptării unor măsuri adecvate pentru atenuarea acestor riscuri, atunci când este cazul, și proporțional cu amploarea riscului.

În baza dezvoltărilor din capitolul I, capitolul II (Realising Trustworthy AI) oferă orientări cu privire la modul în care poate fi realizată o IA de încredere, translatând principiile etice în *șapte cerințe* cheie pe care ar trebui să le îndeplinească sistemele de IA, precum și metodele tehnice, cât și cele non-tehnice ce pot fi utilizate pentru punerea lor în aplicare:

- să se asigure că dezvoltarea, implementarea și utilizarea sistemelor de IA îndeplinesc cele șapte cerințe-cheie pentru o IA de încredere:
  1. agenția și supravegherea umană,
  2. robustețea și siguranța tehnică,
  3. confidențialitatea și guvernanța datelor,
  4. transparența,
  5. diversitatea, nediscriminarea și echitatea,
  6. bunăstarea societății și a mediului
  7. responsabilitatea.
- Analizarea metodelor tehnice și non-tehnice necesare pentru a asigura punerea în aplicare a acestor cerințe;
- Stimularea cercetării și inovării pentru a contribui la evaluarea sistemelor de IA și pentru a promova îndeplinirea cerințelor; diseminarea rezultatelor și a întrebărilor deschide către publicul larg și formarea în mod sistematic a unei noi generații de experți în etica IA;

[1] Toate declarațiile normative din acest document au scopul de a reflecta orientările în vederea realizării celei de-a doua și a celei de-a treia componente a IA de încredere (IA etică și robustă). Adresat tuturor părților interesate, documentul urmărește să meargă dincolo de o listă de principii etice, oferind orientări cu privire la modul în care aceste principii pot fi operaționalizate în sistemele socio-tehnice.

- Comunicarea, într-un mod clar și proactiv, de informații către părțile interesate cu privire la capacitățile și limitările sistemului IA, pentru a le permite să aibă așteptări realiste, precum și cu privire la modul în care sunt puse în aplicare cerințele; asigurarea transparenței cu privire la faptul că au de-a face cu un sistem IA;
- Asigurarea trasabilității și posibilității de auditare a sistemelor IA, în special în contexte sau situații critice;
- Implicarea părților interesate de-a lungul întregului ciclu de viață al sistemului IA; stimularea formării și a educației, astfel încât toate părțile interesate să fie conștiente și instruite în domeniul IA de încredere;
- Conștientizarea faptului că ar putea exista tensiuni fundamentale între diferite principii și cerințe; Identificarea, evaluarea, documentarea și comunicarea continuă a compromisurilor și soluțiile găsite în acest sens.

Capitolul III (Assessing Trustworthy AI) prezintă o lista concretă și neexhaustivă de evaluare pentru o IA de încredere, cu scopul de a operaționaliza cerințele stabilite în capitolul II, oferind astfel practicienilor un ghid practic; Această listă de evaluare va trebui adaptată utilizării specifice a sistemului IA:

- Adoptarea unei liste de evaluare a IA de încredere atunci când se dezvoltă, implementează sau utilizează sisteme IA și adaptarea la cazul specific de utilizare în care se aplică sistemul;
- Luarea în considerare a faptului că o astfel de listă de evaluare nu va fi niciodată exhaustivă; asigurarea unei IA de încredere nu înseamnă bifarea unor casuțe, ci identificarea și implementarea continuă a cerințelor, evaluarea soluțiilor și asigurarea unor rezultate îmbunătățite pe tot parcursul ciclului de viață al sistemului IA cu implicarea părților interesate în acest sens.

O secțiune finală a documentului urmărește să concretizeze unele dintre aspectele abordate în întregul cadru, oferind exemple de oportunități benefice care ar trebui urmărite, precum și preocupări critice ridicate de sistemele de IA care ar trebui analizate cu atenție. Deși prezentele orientări urmăresc să ofere îndrumare pentru aplicațiile IA în general, prin construirea unei platforme orizontale pentru a obține o IA de încredere, situații diferite ridică provocări diferite. Prin urmare, ar trebui analizat dacă, pe lângă acest cadru orizontal, este necesară o abordare sectorială, având în vedere specificitatea contextuală a sistemelor IA.

Prezentele orientări nu intenționează să înlocuiască nicio formă de elaborare sau reglementare a politicilor sau reglementărilor actuale sau viitoare și nici nu vizează descurajarea introducerii acestora. Acestea ar trebui să fie văzute ca un document viu care urmează să fie revizuite și actualizate în timp, pentru a asigura relevanța lor continuă pe măsură ce tehnologia, mediile noastre sociale și cunoștințele noastre evoluează. Acest document reprezintă un punct de plecare pentru discuția despre "O IA de încredere pentru Europa".

În afara contextului european (UE), orientările vizează, de asemenea, să stimuleze cercetarea, reflecția și discuțiile privind un cadru etic pentru sistemele IA la nivel mondial.

Prezentăm în Tabelul 3.1 un set de ghiduri care se referă la IA etică, IA responsabilă, IA de încredere. Aceste ghiduri pot fi puncte de pornire pentru operaționalizarea inteligenței artificiale etice și în implementarea procedurilor pentru auditarea și certificarea aplicațiilor IA etice.

Ghid	**Organism**	**Țara**	**Tip organizație**	**An**
Artificial Intelligence. Australia's Ethics Framework: A Discussion Paper	Department of Industry Innovation and Science	Australia	Administrație	2019
Montreal Declaration: Responsible AI	Univ. Montreal	Canada	Universitate	2017

Work in the Age of Artificial Intelligence. Four Perspectives on the Economy, Employment, Skills and Ethics	Ministerul Economiei și al Forței de Muncă	Finlanda	Administrație	2018
AI Guidelines	Deutsche Telekom	Germania	Industrie	2018
How Can Humans Keep the Upper Hand? Report on the Ethical Matters Raised by AI Algorithms	French Data Protection Authority (CNIL)	Franța	Administrație	2019
The Japanese Society for Artificial Intelligence Ethical Guidelines	Japanese Society for Artificial Intelligence	Japonia	Administrație	2018
Sony Group AI Ethics Guidelines	Sony	Japonia	Industrie	2018
Dutch Artificial Intelligence Manifesto	Special Interest Group on Artificial Intelligence (SIGAI), ICT Platform Netherlands	Olanda	Administrație	2018
Artificial Intelligence and Privacy	The Norwegian Data Protection Authority Norway	Norvegia	Administrație	2018
Mid- to Long-Term Master Plan in Preparation for the Intelligent Information Society	Government of the Republic of Korea	Coreea de Sud	Administrație	2017
DeepMind Ethics & Society Principles	DeepMind Ethics & Society	UK	Industrie	2017
The Responsible AI Framework	Price water house Coopers UK	UK	Industrie	N.A.
Responsible AI and Robotics. An Ethical Framework	Accenture UK	UK	Industrie	N.A.
The AI Now Report 2016. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA	Institut	2016
The AI Now Report 2017. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA	Institut	2017
The AI Now Report 2018. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA	Institut	2018
The AI Now Report 2019. The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term	AI Now Institute	USA	Institut	2019

Statement on Algorithmic Transparency and Accountability	Association for Computing Machinery (ACM)	USA	Asociație	2017
AI—Our Approach	Microsoft	USA	Industrie	N.A.
AI Principles	Future of Life Institute	USA	Institut	2017
IBM's Principles for Trust and Transparency	IBM	USA	Industrie	2018
OpenAI Charter	OpenAI	USA	Industrie	2018
Our Principles	Google	USA	Industrie	N.A.
AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations	AI4People	EU	Asociație	2018
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence	EU	Grup de lucru	2019
Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems	European Group on Ethics in EU Science and New Technologies	EU	Grup de lucru	2018
Top 10 Principles for Ethical Artificial Intelligence	UNI Global Union	Internațional	Asociație	2018
The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation	Future of Humanity Institute; University of Oxford; International Centre for the Study of Existential Risk; University of Cambridge; Center for a New American Security; Electronic Frontier Foundation; OpenAI	Internațional	Asociație	2018
The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems	Access Now; Amnesty International	Internațional	Asociație	2018
White Paper: How to Prevent Discriminatory Outcomes in Machine Learning	WEF, Global Future Council on Human Rights International 2016-2018	Internațional	Grup de lucru	2018
Privacy and Freedom of Expression in the Age of Artificial Intelligence Privacy International & Article 19	Internațional	ONG	2018	

Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Fairness, Accountability, and Transparency in Machine Learning (FATML)	Internațional Grup de lucru	2018

Tabela 3.1: Ghiduri pentru inteligența artificială etică

Jurnale și cărți pentru etică

- AI and Ethics (Springer)
- The AI Ethics Journal (AIRES)
- Artificial Intelligence - Law, Policy, & Ethics (SSRN)
- Ethics of AI in Context: A Multidisciplinary & Multimedia Journal(C4E)

Institute și organizații cu contribuții în IA etică

Institute asociate universităților

- AI Now Institute at New York University
- Institute for Ethics in AI at Oxford
- Institute for Ethics in AI at TUM
- DataLab at AARHUS

Institute independente

- Data & Society
- Partnership on AI
- The Institute for AI and Ethical ML
- Montreal AI Ethics Institute
- ALLAI
- British Centre for Data Ethics and Innovation

Grupuri de lucru

- IEEE Algorithmic Bias WG

Altele

- AI4EU Observatory on Society and Artificial Intelligence (OSAI)

3.3 Evaluarea IA de încredere

A fost elaborată o "Listă de evaluare pentru IA de încredere" pentru a determina în ce măsură o aplicație îndeplinește cerințele. Orientările AI-HLEG pot fi considerate o directivă etică primară pentru dezvoltarea unor sisteme IA de încredere, datorită gândirii și expertizei care au dus la crearea acesteia și a sprijinului Comisiei Europene pentru o abordare centrată pe om a IA.

Al treilea capitol al acestor orientări conținea o listă de evaluare pentru a evalua dacă sistemul IA care este dezvoltat, implementat, achiziționat sau utilizat respectă cele șapte cerințe cheie ale IA de încredere, specificate în document. Această listă de evaluare pentru IA de încredere (ALTAI) este destinată pentru *scop de autoevaluare*. Aceasta oferă o abordare inițială pentru evaluarea IA de încredere; ea se bazează pe Orientările de etică pentru o IA de încredere și a fost dezvoltat pe o perioadă de doi ani, din iunie 2018 până în iunie 2020. În această perioadă, ALTAI a beneficiat, de asemenea, de o fază-pilot (a doua jumătate a anului 2019). Prin această etapă-pilot, AI HLEG a primit feedback valoros prin intermediul a cincizeci de interviuri aprofundate cu companii selectate, printr-un flux de lucru deschis cu Alianța IA pentru a oferi cele mai bune practici și, prin intermediul a două chestionare accesibile public pentru părțile interesate tehnice și non-tehnice.

ALTAI se bazează în mod ferm pe protecția drepturilor fundamentale ale cetățenilor, care este termenul utilizat în Uniunea Europeană pentru a face referire la drepturile omului consacrate în tratatele UE, Carta drepturilor fundamentale (Carta) și Dreptul internațional al drepturilor omului.

ALTAI este *destinată utilizării flexibile*: organizațiile se pot baza pe elemente relevante pentru sistemul IA specific din prezenta listă de evaluare pentru IA de încredere, sau pot adăuga elemente la aceasta după cum consideră de cuviință, luând în considerare sectorul în care își desfășoară activitatea. Aceasta ajută organizațiile să înțeleagă ce este IA de încredere, în special ce riscuri ar putea genera un sistem IA și cum să reducă la minimum aceste riscuri, maximizând în același timp beneficiile IA. Scopul său este de a ajuta organizațiile să identifice modul în care sistemele de IA propuse ar putea să genereze riscuri și să identifice dacă și ce fel de măsuri active ar putea fi necesar să fie luate pentru a evita și a reduce la minimum aceste riscuri.

Organizațiile vor obține cea mai mare valoare din această listă de evaluare prin implicarea activă în întrebările pe care le ridică, care au scopul de a încuraja reflecția atentă pentru a provoca acțiuni adecvate și pentru a cultiva o cultură organizațională angajată în dezvoltarea și menținerea unor sisteme de IA de încredere. Aceasta sensibilizează publicul cu privire la impactul potențial al IA asupra societății, mediului, consumatorilor, angajaților și cetățenilor (în special copiii și persoanele care aparțin grupurilor marginalizate), încurajând implicarea tuturor părților interesate relevante. Aceasta ajută la obținerea de informații cu privire la faptul dacă soluțiile sau procesele semnificative și adecvate pentru a realiza respectarea celor șapte cerințe (așa cum s-a subliniat mai sus) sunt deja în vigoare sau trebuie să fie puse în aplicare. Acest lucru ar putea fi realizat prin orientări interne, procese de guvernanță etc.

O abordare demnă de încredere este esențială pentru a permite "competitivitatea responsabilă", oferind baza pe care toți cei care utilizează sau sunt afectați de sistemele IA pot avea încredere că proiectarea, dezvoltarea și utilizarea lor sunt legale, etice și solide.

ALTAI contribuie la *promovarea inovării responsabile și durabile în domeniul IA în Europa*. Acesta urmărește să facă din etică un pilon de bază pentru dezvoltarea unei abordări unice a IA, una care urmărește să aducă beneficii, să capaciteze și să protejeze atât înflorirea umană individuală, cât și binele comun al societății; acest lucru va permite Europei și organizațiilor europene să se poziționeze ca lideri mondiali în domeniul IA de vârf, demnă de încrederea noastră individuală și colectivă.

Cartea albă privind inteligența artificială

Ca un pas spre asigurarea conformității cu aceste directive, în mod concentric prin legislația viitoare, CE a emis o carte albă privind inteligența artificială pe 19 februarie 2020. În această carte albă, CE prezintă propuneri pentru promovarea dezvoltării IA în Europa, asigurând în același timp respectarea drepturilor fundamentale ale omului. O parte importantă a acestei cărți albe este propunerea de a crea o *evaluare prealabilă a conformității* pentru aplicațiile IA cu risc ridicat, pe baza orientărilor în domeniul eticii ale AI-HLEG. Acest cadru juridic ar trebui să abordeze riscurile pentru drepturile fundamentale și siguranță.

Acest cadru juridic ar trebui să abordeze principalele riscuri legate de utilizarea IA și se referă la aplicarea unor norme menite să protejeze drepturile fundamentale (inclusiv datele cu caracter personal și protecția vieții private și nediscriminarea), precum și siguranța (inclusiv aspectele de securitate cibernetică), dar și aspectele legate de răspundere.

Concluzia prezentată în Cartea albă este că abordarea europeană pentru IA urmărește să *promoveze capacitatea de inovare a Europei în domeniul IA*, sprijinind în același timp *dezvoltarea și adoptarea IA etice și de încredere în întreaga economie a UE*. IA ar trebui să funcționeze pentru oameni și să fie o forță a binelui în societate.

3.4 Principii etice pentru IA

Pe linia propusă de The Institute for AI and Ethical ML, cadrul practic pentru a dezvolta responsabil IA se bazează pe opt principii:

Augmentare umană: înțelegerea consecințelor predicțiilor incorecte, mai ales atunci când se automatizează procese critice (de exemplu, în justiție, sănătate, transport etc.); permiterea intervenției umane în cadrul sistemelor ML.

Evaluarea bias-ului: focus asupra dezvoltării de procese și metode pentru identificarea și documentarea bias-ului inerent în date, trăsături și rezultatele predicțiilor, precum și înțelegerea implicațiilor acestor tipuri de bias, astfel încât procesele potrivite să poată fi puse în aplicare pentru a atenua potențialele riscuri. Implementarea unui sistem biased poate avea ca efect consolidarea unor nedreptăți sociale existente. Aspectele etice ar trebui luate în considerare împreună cu părțile interesate relevante din industrie (comitete etice, organisme de reglementare etc.)

Interpretabilitate prin justificare: îmbunătățirea transparenței și a interpretabilității modelelor de învățare automată acolo unde este rezonabil (efort continuu de a îmbunătăți instrumentele și procesele care permit explicarea rezultatele pe baza caracteristicilor și modelelor alese; adăugarea de cunoștințe de domeniu în loc de învățare de trăsături latente, chiar cu prețul unei performanțe mai scăzute).

Operații reproductibile: nivel rezonabil de reproductibilitate a operațiilor, prin abstractizarea grafurilor de calcul și arhivarea datelor la fiecare etapă a pașilor de procesare (posibilitatea de revenire la versiunea anterioară a unui model, reproducerea unei intrări pentru a depana o funcționalitate specifică etc.); adoptarea de standarde deschise; abstractizarea pașilor de procesare.

Strategie de reorientare profesională: atenuarea impactului la nevelul societății produs de automatizarea sarcinilor ce revin unor meserii specifice; susținerea părților interesate să dezvolte o strategie de gestionare a schimbărilor atunci când implementează tehnologia nouă.

Acuratețe practică: corelarea parametrilor de optimizare cu perspectivele specifice domeniului; ceea ce poate fi „corect" pentru un computer, poate fi „greșit" pentru un om (și invers); transcenderea dincolo de acuratețe; tehnologiile ar trebui să înțeleagă și să aplice elementele fundamentale în orice moment și să se asigure că analizează implicațiile diferitelor tipuri de erori, precum și care ar trebui să fie modul corect de evaluare a acestor erori în contextul specific domeniului (de exemplu, să evalueze impactul diferitelor tipuri de erori).

Încredere prin confidențialitate: asigurarea confidențialității distribuite la nivel de proiectare, precum și punerea la punct de procese continue pentru a construi încredere nu numai cu utilizatorii, ci și cu părțile interesate relevante, precum cadrele de achiziții, utilizatorii operaționali, și nu numai.

Confidențialitate la nivelurile potrivite:] utilizarea de către Uber a confidențialității diferențiale este un exemplu, în care s-a introdus un sistem care adaugă zgomot la rezultatele interogării, în care zgomotul este relativ la nivelul de granularitate cerut de interogare, pentru a se asigura că analiza continuă să aibă acces la seturile de date relevante, evitând în același timp expunerea informațiilor personale prin metadate: decideți dacă metadatele pot expune informații personale (a se vedea scandalul Cambridge Analytica)

Conștientizarea riscului datelor: sistemele autonome de luare a deciziilor deschid porțile către noi potențiale încălcări ale securității; un procent mare de încălcări ale securității apar din cauza erorilor umane, spre deosebire de hacks-urile reale; luarea de măsuri explicite, cum ar fi educarea personalului relevant, stabilirea de procese clare în jurul datelor și evaluarea implicațiilor atacurilor asupra modelelor ML (cum ar fi atacurile adversariale)

Tipuri de bias
Bias de eșantionare - prezent în datele de antrenare
Bias produs de prejudecăți - bazat pe clasa socială, rasa, naționalitate, gen
Bias de confirmare - accepțiunea psihologică, de a valida mult mai ușor ipoteze și rezultate care sunt aliniate cu credințele sau presupunerile designerilor de sisteme ML
Bias de încadrare - rezultă din antrenarea unui model pe date care conțin o perspectivă asimetrică asupra unui anumit grup social.

■ **Exemplu 3.1 — Interpretabilitate prin justificare.** O serie de inițiative și unelte pentru învățare

automată etică sunt disponibile la EhicalML ■

Pentru operaționalizarea celor opt principii s-a propus un set de criterii care asigură existența infrastructurii tehnice de bază și corectitudinea proceselor, cu denumirea Machine Learning Maturity Model:

1. Seturi de date benchmark practice ⇒ Principiul 6: Precizie practică
2. Interpretabilitate prin justificare ⇒ Principiul 3: Interpretabilitate prin justificare
3. Infrastructură pentru operații reproductibile ⇒ Principiul 4: Operații reproductibile
4. Procese de evaluare a datelor și modelelor ⇒ Principiul 2: Evaluarea bias-ului
5. Capabilități de aplicare a confidențialității ⇒ Principiul 7: Încredere prin confidențialitate
6. Proiectarea procesului operațional ⇒ Principiul 1: Augmentare umană
7. Capacități de gestionare a schimbărilor ⇒ Principiul 5: Strategie de reorientare profesională
8. Atenuări ale riscului de securitate ⇒ Principiul 8: Conștientizarea riscului datelor

R UE a adoptat "European Charter for Robots and Humans" in 28 Ianuarie 2021. De asemenea, NU a redactat un draft on “Robotics Ethics“ [49] prin care roboții trebuie să respecte anumite norme etice.

3.5 Abordarea industriei

Facebook - Oversight Board. entitate independentă pentru a susține dreptul oamenilor la libera exprimare și pentru a se asigura că aceste drepturi sunt respectate în mod adecvat. Deciziile consiliului de a susține sau de a inversa deciziile de conținut Facebook vor fi obligatorii, ceea ce înseamnă că Facebook va trebui să le pună în aplicare, cu excepția cazului în care acest lucru ar putea încălca legea.

Google - Ethical AI team Inlocuitor al Ethics Board[a][b]. Princiile enunțate pentru IA etică sunt:

- Cercetare în beneficiul societății
- Evitarea creării sau întăririi prejudecății neloiale
- Siguranța
- Responsabilitate față de oameni
- Asigurarea confidențialității prin proiectare
- Standarde de excelență științifică

[a] Stabilită în aprilie 2019, dizolvată la nici doua săptămani după înființare, https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board

[b] scandal la început de 2021 după ce au fost concediate la o saptamana distanța cele doua coordonatoare ale echipei - Timnit Gebru si Margaret Mitchell, https://venturebeat.com/2021/03/12/ai-weekly-facebook-google-and-the-tension-between-profits-and-fairness/

Axon - AI Ethics Board.

- Orice produs nou este analizat de consiliu
- Construirea de unelte care permit transparența și controlul uman asupra modului în care tehnologiile IA sunt folosite
- Transparența algoritmilor și analiza modului în care aceștia ar putea fi folosiți greșit
- Disponibilitatea datelor pe care modelele ML au fost antrenate, a intrărilor folosite la inferență
- Disponibilitatea măsurilor de securitate a datelor și confidențialitate

DeepMind - Ethics and Society Team. Temele abordate sunt:

- Confidențialitate, transparență și corectitudine
- Moralitatea și valorile IA
- Guvernanță și responsabilitate

- IA și provocările complexe ale lumii
- Utilizarea greșită și consecințele neintenționate
- Impact economic: incluziune și egalitate

Echipa include consultați de la universități precum Oxford, Columbia, McKinsey, Princeton, Cambridge.

Microsoft - FATE: Fairness, Accountability, Transparency, and Ethics in AI Analiza implicațiilor sociale complexe ale IA, învățării automate sau a procesării limbajului natural. Scopul echipei este de a promova tehnicile de calcul care sunt atât inovatoare, cât și responsabile, prioritizând în același timp problemele de echitate, responsabilitate, transparență și etică, în măsura în care acestea sunt legate de IA, ML și NLP, bazându-se pe câmpuri cu orientare socio-tehnică, cum ar fi interacțiunea om-mașină, știința informației, sociologie, antropologie, studii științifice și tehnologice, studii media, științe politice și drept.

Complementar, Social Media Collective are scopul de a oferi o înțelegere contextuală a dinamicii sociale și culturale care stă la baza tehnologiilor de socializare.

Amazon Ethics in AI. Nu pare sa aibă o echipă dedicată pe IA etică, dar există interes spre zona de etică și inițiative în cadrul Amazon precum:

- FATE - Fairness, Accountability, Transparency, Ethics
- AI fairness projects
- Etică și XAI

R EEE P037 Algorithmic Bias WG are termen Decembrie 2021 pentru a propune o metodologie de certificare care să stabilească responsabilitatea și să clarifice modul în care algoritmii targetează, evaluează și influențează utilizatorii și actorii implicați.

3.6 Aplicații suport pentru IA etică

Unelte pentru XAI

Captum: bibliotecă cu unelte de interpretabilitate și înțelegere a modelelor, dezvoltată de Facebook

ELI5 ("Explain it like I'm 5"): - importanța trăsăturilor în modele, interpretabilitate pentru predicțiile pe imagini.

Deep Visualization Toolbox: o colecție de tehnici de vizualizare la nivel de neuron pentru rețele adânci, care încearcă să evidențieze ceea ce "vede" (activează) un neuron.

IBM XAI 360: un set de algoritmi și metrici de explicabilitate (e.g. monotonicitatea și fidelitatea), care conține atât modele globale, cât și locale, directe cât și post-hoc;

Suita Tensorflow: WhatIf, Cleverhans, Lucid, Model Analysis

eXplainableAI: bibliotecă pusă la dispoziție de EthicalML Institute, ce conține unelte pentru evaluarea datelor și a modelelor obținute prin învățarea automată.

Unelte pentru evitarea biasului

IBM AI Fairness 360 : conține un set de metrici pentru evaluarea biasului în date și modele, explicații pentru aceste metrici, dar și algoritmi pentru evitarea sau corectarea biasului;

Global Explanations for Bias Identification (GEBI): descoperirea biasului cu explicatii globale, infuzie voluntară de bias și verificarea impactului asupra modelului de predicție

Responsibly : unealtă "one-shop-stop" pentru auditarea biasului și echitabilității în învățarea automată și corectarea acestuia prin intervenții algoritmice; focus pe modele de procesare a limbajului natural.

4. Reglementarea IA în sectoare de activitate

Rezumat. IA va afecta toate sectoarele de activitate. O prioritizare a acestora cu scopul de a facilita dezvoltarea IA pentru un domeniu specific ar fi o decizie bazată pe multe variabile dinamice în contextul dezvoltării extrem de rapide a tehnologiilor din IA. De aceea, o abordare în care *pătrunderea sistemelor IA este sprijinită echidistant în toate sectoarele* poate fi o soluție corectă. Principalul criteriu pentru decizii curente legate de finanțarea proiectelor IA ar putea fi *valoarea adăugată adusă de IA în domeniul sau scenariul respectiv*. La nivel UE, trei domenii în care IA aduce cea mai mare valoare adăugată ar putea fi: (i) canale de distribuție, (ii) mobilitate și orașe inteligente, (iii) sănătate.

Dezvoltarea IA în România ține în primul rând de dezvoltarea sectorului IT, de expertiza și dimensiunile departamentelor de cercetare și dezvoltare ale companiilor sau institutelor de cercetare. Un aspect relevant în stimularea IA pentru anumite sectoare este și dimensiunea pieței de IA în România. În ipoteza unei piețe de vânzare mici la nivel național pentru sectoare specifice, dezvoltarea IA va fi influențată în primul rând de domeniile care vor cere soluții IA la nivel european sau global[a].

Exemplificăm în acest capitol o serie de aplicații IA în diferite sectoare de activitate și facem referiri la posibila clasificare a acestor aplicații din perspectiva grupelor de risc propuse de AIA. De asemenea, exemplificăm o parte din reglementările în vigoare, în contextul în care AIA este un cadru de reglementare pentru inteligența artificială care se bazează pe standarde și pe utilizarea pe cât posibil a experienței și reglementărilor din sectoarele specifice.

[a] O abordare complementară celei de prioritizare a sectoarelor ar fi raportarea soluțiilor IA la țelurile de dezvoltare sustenabilă (Sustainable Development Goals - SDG). Agenda pentru dezvoltare sustenabilă pentru 2030 adoptată de către UN în 2015 stabilește 17 obiective: (i) eliminarea sărăciei; (ii) zero foame; (iii) sănătate; (iv) educație de calitate; (v) egalitatea genurilor; (vi) apă curată; (vii) energie curată și accesibilă; (viii) muncă decentă și creștere economică; (ix) industrie, inovatie și infrastructură; (x) reducerea inegalităților; (xi) orașe și comunități sustenabile; (xii) consum și producție responsabile; (xiii) acțiuni climatice; (xiv) viața sub apă; (xv) viața pe uscat; (xvi) pace, justiție și instituții puternice; (xvii) parteneriate pentru aceste obiective.

4.1 Producție și canale de distribuție

A patra revoluție industrială are ca elemente centrale nu doar un model ”digital-first“, ci și utilizarea de sisteme IA pe întreg lanțul de producție. Deși doar 9% din companiile din domeniul producției utilizează sisteme de IA în mod curent (conform PWC), impactul pe care IA îl poate avea asupra

utilajelor, oamenilor și proceselor de producție va fi semnificativ pe termen mediu și lung[1]:

Mentenanța echipamentelor: sistemele de mentenanță predictivă cu sau fără senzori ce transmit informație în timp real pot determina prin modele IA predictive când trebuie înlocuite utilaje sau componente, făcând producția mai puțin susceptibilă la întreruperi. Această abordare proactivă a mentenanței poate duce la reducerea cu până la 75%[2] a întreruperilor cauzate de defecțiuni sau revizii

Managementul clădirilor și siguranța la locul de muncă: prin folosirea de camere de supraveghere avansate, capabile să determine evenimente, să identifice persoane, combinate cu sisteme IA capabile să stabilească răspunsuri automate la situații de risc

Securitate cibernetică: pe măsură ce numărul de echipamente conectate într-o unitate de producție crește, pe atât crește și riscul de atacuri cibernetice. De exemplu, pentru detectarea anomaliilor în rețelele informatice se folosesc în mod curent tehnici de învațare automată. De asemenea, sunt dezvoltate sisteme de IA care pot contracara, in mod automat, atacuri informatice

Asigurarea eficienței și calității în producție: prin folosirea de modele specifice bazate pe viziune artificială pentru sisteme de stocare și transport automate (e.g. paletizare, ambalare), managementul flotei de utilaje de producție și transport, detectarea de defecte pe linia de producție și asigurarea calității.

Domeniul industrial și de producție încorporează un spectru foarte larg de aplicații IA, de la cele cu risc minim (e.g. mentenanță predictivă) până la cele cu risc ridicat (e.g. dispozitive autonome). Reglementările IA pot vor avea un impact semnificativ asupra acestui domeniu, în sensul standardizării și creșterii încrederii în sistemele IA.

4.2 Mobilitate și orașe inteligente

IA se regăsește deja în domeniul transportului, cu previziuni de impact decisiv în următorii ani. Dincolo de inițiativele de a introduce vehicule autonome[3], printre elementele cu impact semnificativ legate de rețeaua de transport, subliniem:

Managementul traficului: reducerea congestiilor și blocajelor din trafic folosind modele de IA predictive, sau folosind senzori și tehnologia IoT pentru raportare în timp real, inclusiv modele care pot determina zone de risc ale accidentelor. Pentru transportul aerian, modele de IA care îmbunătățesc rutele și consumul de carburant sau alte resurse.

Eficientizarea călătoriilor aeriene: una dintre cele mai costisitoare aspecte ale călătoriei cu avionul pentru populație este întârzierea zborurilor, jumătate din pierderi fiind suportate financiar de călători[4].

Navigație, porturi: logistica transportului maritim poate beneficia substanțial de soluții IA, de optimizarea proceselor interne cu AI-enabled Robotic Process Automation (RPA), la utilaje inteligente, urmărirea mărfurilor cu IoT, optimizarea rutelor și folosirea de modele predictive[5].

Siguranța transportului. Siguranța pasagerilor și a mărfurilor este deja îmbunătățită cu ajutorul sistemelor de suport implementate la vehicule, de la sisteme de menținere a distanței față de vehiculele din proximitate, navigație cu control adaptiv, frânare automată sau detectare de obstacole. Chiar dacă multe din aceste sisteme nu folosesc în mod curent IA, noile versiuni

[1] https://www.themanufacturer.com/articles/ai-transforming-manufacturing

[2] http://www.moorinsightsstrategy.com/wp-content/uploads/2016/08/IoT-Analytics-at-the-Edge-by-Moor-Insights-and-Strategy.pdf

[3] https://www.forbes.com/sites/cognitiveworld/2019/07/26/how-ai-can-transform-the-transportation-industry/?sh=26e0abbd4964

[4] https://news.berkeley.edu/2010/10/18/flight_delays

[5] https://amconsoft.com/how-ai-is-transforming-the-transportation-industry/#9_Artificial_intelligence_in_shipping_navigation_and_ports

utilizează viziunea artificială pe bază de învățare automată. Aceste sisteme de asistență a șoferului sunt deja supuse unor norme europene de siguranță. Un exemplu in acest sens ar fi controlul adaptiv al navigației [54].

R Atât sistemele de siguranță și asistență pentru șofer, cât și vehiculele autonome care folosesc IA prezintă o provocare din perspectiva reglementării. Potrivit reglementării IA propuse la nivel european, aceste sisteme se încadrează ca IA de risc ridicat.

Vehicule autonome. Este unul dintre primele domenii ale inteligenței artificiale care a necesitat reglementare. În 2016 Departamentul pentru Transport al SUA a emis Federal Automated Vehicles Policy [102]. Sunt abordate subiecte precum: condiții de test, cerințe pentru vehiculul autonom, condiții de asigurare.. În februarie 2017, Vehicle Technology and Aviation Bill din Regatul Unit clarifică in domeniul asigurărilor aspecte legate de compensațiile oferite victimelor accidentelor cu vehicule autonome. Șoferul nu este exclus din procesul de determinare a vinovăției.

4.3 Sănătate

Domeniul sănătății poate fi eficientizat pe de o parte prin utilizarea IA în gestiunea resurselor, iar pe de altă parte din perspectiva actului medical de la diagostic, la medicație și îngrijirea pacienților, activități în care expertiza umană specifică poate beneficia de unelte suport bazate pe IA.

Deși aplicațiile curente ale IA în domeniul medical sunt fascinante din punct de vedere al metricilor de perfomanță raportate, există și studii care scot în evidență dificultățile

■ **Exemplu 4.1 — Punct de vedere.** „Sistemele de IA nu sunt suficient de specifice pentru a înlocui dubla citire radiologică în programele de screening", se argumentează într-o cercetare din 2021[6]. S-au analizat 12 studii care au raportat acuratețea testării radiologice cu algoritmi IA în comparație cu radiologi umani. În cele mai mari trei studii retrospective, care au inclus 79.910 femei examinate în Europa și Statele Unite, majoritatea sistemelor de IA (34 din 36 de sisteme, 94%) au fost mai puțin precise decât un singur radiolog și toate au fost mai puțin exacte decât screeningul de către doi sau mai mulți radiologi. În plus, rezultatele promițătoare din studiile mai mici nu au fost reproduse în cele mai mari și nu au existat studii prospective din lumea reală care să măsoare acuratețea testului cu IA în aceste studii mai mici. Un alt studiu al Stanford[7] arată că multe modele de IA nu sunt documentate cu rigoarea sau transparența pe care profesioniștii din domeniul medical le consideră necesare, indicând un motiv important pentru care modelele de IA au performanțe surprinzător de slabe în condiții medicale reale, chiar și după ce au obținut rezultate bune în faza de laborator. ■

R Companiile din domeniul sănătății operează într-un mediu cu multe reglementări specifice. Ca urmare, ciclul de trecere de la idee la produs rămâne mare (e.g. medicament, dispozitiv medical), chiar dacă IA ajută la micșorarea lui. Schemele de ajutor financiar pentru sprijinirea IA în domeniul medical ar trebui să depindă de durata ciclului de trecere de la idee la produs.

Partajarea datelor în domeniul medical) rămâne un subiect sensibil. O linie posibilă se referă la utilizarea de tehnogii IA care permit învățarea automată fără a partaja datele. O astfel de soluție tehnică ar putea fi învățarea federată (i.e. Federated learning).

Dezvoltarea de soluții IA în domeniul medical necesită o foarte mare atenție, riscul fiind ridicat nu doar din cauza potențialei lipse de acuratețe, dar și din cauza tendinței specialiștilor de a acorda încredere sistemelor, fără a verifica sau contesta pe termen lung calitatea rezultatelor acestora.

[6] https://emedicine.medscape.com/article/1945498-overview
[7] https://hai.stanford.edu/news/flying-dark-hospital-ai-tools-arent-well-documented

4.4 Finanțe și bănci

Domeniul financiar și bancar utilizează inteligența artificială cu succes, în aplicații precum învățarea automată, procesare naturală de limbaj sau viziune artificială. De exemplu, Accenture raportează că „băncile pot realiza o creștere de 2-5 ori a volumului de interacțiuni sau tranzacții cu același număr de angajați" utilizând instrumente bazate pe IA[8]. IA este folosit de bănci și instituții financiare[9] pentru:

Reducerea costurilor operaționale și a riscului: Procesele bancare sunt dependente de fluxul de documente interne, care poate fi eficientizat prin implementarea de Robotic Process Automation (RPA) în combinație cu IA, pentru a automatiza procese repetitive și a realiza operațiuni cognitive simple cu ajutorul procesării de limbaj natural (i.e. NLP), subdomeniu cunoscut ca Intelligent Document Processing (IDP).

Îmbunătățirea experienței clienților: prin intermediul chatbots sau sisteme intelligente care pot interpreta și ruta mesaje de pe mai multe canale (scris/email, voce, video) către agenți umani sau software. Self-service, conceptul prin care clienții pot să se ajute singuri în cazul unor probleme, este suportat de IA prin implementarea de sisteme automate de clasificare, extragere intenții și cuvinte cheie, dar și modele IA conversaționale.

Îmbunătățirea detectării fraudei și a respectării reglementărilor: Detectarea fraudelor dar și respectarea regulamentelor exhaustive la care sunt supuse instituțiile financiare se face și în prezent cu utilizarea IA, de la monitorizarea anomaliilor în procese și rețele, la monitorizare comunicațiilor multi-channel și detectarea automată de acțiuni care pot constitui fraude sau încalcă legea.

Îmbunătățirea deciziilor privind împrumuturile și creditele: Cea mai controversată utilizare a IA în domeniul bancar este analiza automată a scorului financiar al clienților, pe baza căruia se acordă credite și împrumuturi. Cu toate că calcularea unui asemenea scor se face încă și manual de către bănci și constituie o practică internă legală și acceptată, utilizarea unor sisteme de IA care incorporează date de antrenare irelevante, ne-etice sau cu bias discriminatoriu, poate rezulta în decizii automate discutabile. Rezultatul, în loc să fie o automatizare a unui proces manual laborios dar repetitiv și o durată mai scurtă de decizie, poate să fie discriminarea prin algoritmi [10].

Automatizarea procesului de investiții: Firme ca UBS, cu sediul în Elveția și ING, cu sediul în Olanda, au sisteme de IA care monitorizează piețele de investiții pentru oportunități de investiții neexploatate și își informează sistemele de tranzacționare algoritmice. În timp ce oamenii sunt încă la curent cu toate aceste decizii de investiții, sistemele de IA descoperă oportunități suplimentare printr-o mai bună modelare și descoperire.

Referitor la reglementările de IA care au impact în domeniul financiar bancar, se observă că acest sector poate cuprinde tot spectrul de IA, de la cel cu risc foarte scăzut (e.g. predicții de investiții pe stocuri), până la IA cu risc ridicat (e.g. scor pentru creditare).

4.5 Justiție

Utilizarea IA în justiție ar putea ameliora analiza și colectarea de date, precum și protejarea victimelor. Acest aspect ar putea fi investigat și însoțit de evaluări ale impactului, în special în ceea ce privește garanțiile pentru un proces echitabil, precum și împotriva părtinirii și a discriminării.

IA oferă judecătorilor instrumente pentru a stabili moduri de echilibrare a cauțiunii între

[8] https://www.accenture.com/_acnmedia/PDF-138/Accenture-Banking-AI.pdf

[9] https://searchenterpriseai.techtarget.com/feature/AI-in-banking-industry-brings-operational-improvements

[10] https://factsndata.com/algorithm-Bias-a-prominent-trend-in-credit-score-ratings-impacting-banks-and-consumers.alike.php

drepturile inculpaților și nevoia de siguranță publică. Unele decizii sunt deja luate pe baza uor sisteme de evaluare a riscului, care se doresc a fi cât mai transparente [11].

Natura independentă a sistemelor IA poate ridica un grad înalt de incertitudine și de risc[12] fiind nevoie de o înțelegere clară și detaliată a tuturor mijloacelor prin care inteligența artificială ar putea afecta profesiile și întreaga funcționare a justiției.

Utilizarea sistemelor de IA în procesele de luare a deciziilor judiciare, prin permiterea unor rezultate judiciare programabile și deseori previzibile, atrage o mulțime de provocări și riscuri semnificative pentru dreptul la proces echitabil și pentru administrarea justiției.

Utilizarea IA în justiție vine și din nevoia de a crește nivelul de furnizare a justiției, de a face justiția mai accesibilă, mai rapidă și mai puțin costisitoare, prin urmare gradul de mulțumire al cetățenilor va crește accentuat. Deoarece pentru o parte din populație accesul la o justiție corectă rămâne un lux inaccesibil, posibilitatea de judecare cu ajutorul unei mașini poate fi considerată un progres semnificativ. Utilizarea inteligenței artificiale ca instrument de luare a deciziilor ar putea permite judecătorilor să facă acțiuni de judecată mai coerente și de calitate superioară, eficient și intr-un mod rațional.

În ceea ce privește gradul de risc a sistemelor IA dedicate administrării justiției, alineatul (40) din noul regulament al UE privind inteligența artificială [28] stabilește că

> "Anumite sisteme de IA destinate administrării justiției și proceselor democratice ar trebui clasificate ca având un grad ridicat de risc, având în vedere impactul potențial semnificativ al acestora asupra democrației, statului de drept și libertăților individuale, precum și asupra dreptului la o cale de atac eficientă și la un proces echitabil."

Totodată, în Anexa 3, punctul 8, se sublinează că sistemele de IA cu grad ridicat de risc conform articolului 6 alineatul (2) sunt sistemele de IA enumerate în oricare dintre următoarele domenii: (i) administrarea justiției și procesele democratice: (ii) sisteme de IA menite să ajute o autoritate judiciară în cercetarea și interpretarea faptelor și a legii, precum și în aplicarea legii la un set concret de fapte.

R Alte programe UE privind integrarea IA în sistemul judiciar: Regulamentul (UE) 2021/694 al Parlamentului European și al Consiliului din 29 aprilie 2021 de instituire a programului „Europa digitală" și de abrogare a Deciziei (UE) 2015/2240, text cu relevanță pentru SEE [29].

Provocarile conexe sunt legate de: (i) asigurarea cooperării transfrontaliere, (ii) îmbunătățirea accesului la justiție pentru cetățeni, întreprinderi, practicienii în domeniul dreptului și membrii sistemului judiciar, prin furnizarea unor interconexiuni având interoperabilitate semantică cu baze de date, (iii) facilitarea soluționării extrajudiciare online a litigiilor. (iv) promovarea dezvoltării tehnologiilor inovatoare pentru instanțe și pentru profesia juridică, care se bazează, printre altele pe soluții care au la bază IA și care sunt în măsură să modernizeze și să accelereze procedurile (de exemplu aplicații de „tehnologie juridică").

R Consiliul Barourilor Europene (Council of Bars and Law Societies of Europe) propune o hartă a posibilelor utilizări ale sistemelor IA în diferite etape ale unei proceduri judiciare (vezi link).

[11] https://engineering.stanford.edu/magazine/article/can-ai-help-judges-make-bail-system-fairer-and-safer

[12] Considerații cu privire la aspectele legale în ceea ce privește Inteligența Artificială disponibil la https://www.unbr.ro/wp-content/uploads/2020/05/RO_07a_Draft-CCBE-considerations-on-legal-aspects-of-AI.pdf

4.6 Agricultura

Potrivit World Economic Forum inovația în agricultură folosind IA se poate dezvolta pe patru paliere:

Planificarea inteligentă a culturilor: utilizarea de modele IA pentru recomandarea de perioade de însămânțare, întreținere și cultivare, planificări la nivel micro și macro, mapări de resurse și evenimente.

Agricultură Smart: implementarea conceptului de Farming-as-a-Service, de la sisteme de irigații smart ce folosesc tehnologii IoT și IA, planificarea rezervelor și utilizarea apei cu modele predictive IA, asigurări ale recoltelor.

De la poarta fermei în farfurie: măsurarea calității produselor folosind IA, trasabilitatea producției, legături inteligente producător-cumpărător, depozitare inteligentă cu IoT și IA.

Agricultură Data-Driven: crearea și gestionarea de dataset-uri relevante pentru agricultură, referitor la evidența sănătății solului, randamentele culturilor, vremea, depozitare, înregistrări funciare, piețe și imagini de dăunători.

Alte exemple de aplicații IA in agricultură sunt:

Utilizarea eficientă a apei. Crearea planificării și distribuției eficiente a irigațiilor în funcție de tipul de plante, descoperirea defecțiunilor de irigare sau a pierderilor

Identificarea bolilor și a dăunătorilor. Sistemele IA analizează automat imaginile și alertează agentul uman in caz de idenditicare a unor boli sau dăunători

Pulverizarea cu precizie a soluțiilor erbicide. Tehnologiile IA pot prezice condițiile meteorologice, analizează sustenabilitatea culturilor și evaluează fermele pentru prezența bolilor sau dăunătorilor și nutriția slabă a plantelor în ferme cu date precum temperatura, precipitațiile viteza vântului și radiația solară.

Majoritatea acestor aplicații ale IA pentru agricultură se încadrează în zona de risc scăzut, integrate în procesele existente de producție și management și supuse standardelor de securitate și calitate existente. Excepția e făcută de folosirea IA în mod neasistat uman (utilaje autonome sau cele care interacționează direct cu oameni și prezintă un risc de siguranță), sau folosirea IA pentru a integra date despre consumatori în strategii de vânzări, marketing sau producție.

4.7 Administrație

Inițiativa AI Watch a Comisiei Europene care monitorizează dezvoltarea, adoptarea și impactul inteligenței artificiale pentru Europa[13] a lansat în luna martie 2021 un sondaj pentru a înțelege impactul IA în sectorul public european și a crea o foaie de parcurs pentru a susține folosirea IA în administrația publică[14].

Un raport comprehensiv legat de utilizarea IA în sectorul public a fost publicat de Comisia Europeană în 2020: Raportul Science for Policy, AI Watch, Inteligența artificială în serviciile publice Prezentare generală a utilizării și impactului IA în serviciile publice din UE. Acest raport a colectat 230 de inițiative de folosire a IA în sectorul public european, trei dintre ele fiind în România, toate la nivelul administrației locale, unul în domeniul serviciilor publice generale și două în domeniul ordinii publice și siguranței. Același raport grupează și tipurile de IA folosite în administrația publică. Referitor la domeniile de utilizarea a IA în sectorul public, același raport sunt amintite cinci categorii de sarcini guvernamentale:

Punerea în aplicare: aplicarea reglementărilor existente

Cercetare, analiză și monitorizare a reglementărilor: utilizarea IA la procesele de elaborare a politicilor, cum ar fi colectarea, monitorizarea și analiza datelor pentru a spori capacitățile de

[13] https://knowledge4policy.ec.europa.eu/ai-watch_en

[14] https://knowledge4policy.ec.europa.eu/news/survey-artificial-intelligence-use-public-sector_en

luare a deciziilor factorilor de decizie și pentru a lua decizii pe bază de dovezi.

Adjudecare: utilizarea IA pentru a asista la acordarea de beneficii sau drepturi.

Servicii publice și implicare: furnizarea de servicii cetățenilor și întreprinderilor sau facilitarea comunicării și participarea publicului

Management intern: gestionarea organizației interne, cum ar fi resursele umane, achizițiile publice, sistemele TIC sau utilități.

Dificultăți pentru implemenetarea IA în sectorul public

- Calitatea datelor și dificultațile de integrare
- Costurile ridicate cu infrastructura pentru a stoca și procesa datele, respectiv resursa umană specializată pe IA
- Dificultătile de angajare a experților pe inteligență artificială [109]

R Centre de cercetare precum Future Society at Harvard Kennedy School sau Future of Humanity Institute at the University of Oxford sugerează necesitatea politicilor pentru inteligență artificială, inclusiv prin crearea unui forum global de guvernanță și monitorizare.

R Este necesară asigurarea transparenței cu privire la toate aplicațiile IA care sunt utilizate de diferite instituții ale statului.

■ **Exemplu 4.2 — Politici pentru achiziția de IA de încredere în sectorul public în UK**[15]**.** În iunie 2020, guvernul Marii Britanii, în colaborare cu Forumul Economic Mondial (WEF), a publicat Ghidul pentru achizițiile de inteligență artificială IA care oferă guvernului central și altor organisme din sectorul public principii directoare pentru achiziționarea tehnologiei IA. Ele ghidează, de asemenea, provocările care pot apărea în timpul procesului de achiziție. În legătură cu acest proiect, Biroul pentru IA și WEF au creat setul de instrumente AI Procurement in a Box toolkit [**toolkit0202**], care acționează ca un ghid pentru achizițiile de IA din sectorul public.

■

Ghid pentru achizițiile de IA din sectorul public[a]

1. Include achizițiile în cadrul unei strategii pentru adoptarea IA
2. Luarea de decizii într-o echipă diversă multidisciplinară pentru a atenua prejudecățile IA
3. Efectuarea unei evaluari a datelor înainte de a începe procesul de achiziție
4. Evaluarea beneficiilor și a riscurilor IA, inclusiv definirea obiectivului beneficiului public
5. Interacționarea eficientă cu furnizorii de IA de la bun început
6. Stabilirea căii corecte către piață și concentrarea pe provocare și nu pe o soluție specifică
7. Elaborarea unui plan pentru guvernanță și asigurarea informațiilor
8. Evitarea algoritmilor netransparenți (i.e. “black box”) și blocarea furnizorului
9. Concentrarea asupra necesității de a aborda limitările tehnice și etice ale IA în timpul evaluării
10. Luarea în considerare a gestionarii ciclului de viață al sistemului IA.

[a] https://www.gov.uk/government/publications/guidelines-for-ai-procurement

R Conform [66], România se află pe locul 55 din 194 țări în clasamentul care măsoară gradul de pregătire a administrației publice pentru adoptarea și utilizarea sistemelor bazate pe IA. Primele 3 poziții sunt ocupate de Singapore, Marea Britanie și Germania.

[15] https://www.gov.uk/government/publications/guidelines-for-ai-procurement

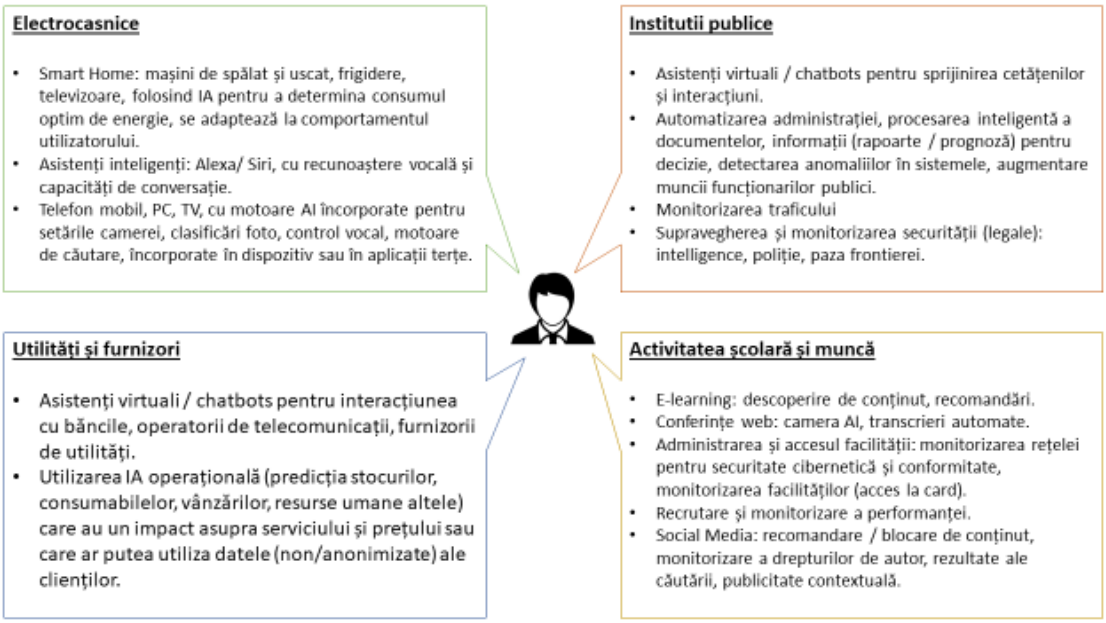

Figura 4.1: Utilizarea IA în diferite domenii

4.8 Utilizare casnică

Interacțiunea utilizatorilor cu sisteme IA se realizează pe mai multe paliere, atât prin interacțiunea cu instituțiile publice, locul de muncă, instituții educaționale, companiile de servicii și utilități, dar și în cadrul personal, casnic (Figura 4.1). Pentru interacțiunile sistemelor IA cu cetățenii, diferite niveluri de reglementare se vor aplica, de la sisteme cu risc scăzut care nu vor necesita reglementare (electrocasnice, asistenți virtuali care nu folosesc date personale și nu efectuează operațiuni de risc), până la sisteme cu risc ridicat (recomandare/blocare continut de știri pe rețele sociale, recrutare și evaluare).

4.9 Educație

Educația este unul dintre domeniile vitale pentru evoluția societății pe termen lung. Ritmul rapid de dezvoltare a tehnologiei predispune oamenii la întârzieri în asimilarea noutăților, atât din perspectivă informațională (i.e. privind cunoașterea și înțelegerea tehnologiei), cât și experiențială (i.e. privind adaptarea prin utilizarea tehnologiei).

Pentru reducerea riscurilor asociate tehnologiei emergente (e.g. erori, accidentări, nedreptăți generate în urma utilizării în lipsă de cunoaștere a tehnologiei, pierderea controlului uman asupra tehnologiei) sunt necesare:

- Derularea unei analize formale și documentate pentru identificarea nevoilor de formare a populației, cu privire la IA (și, în general, la tehnologiile emergente), în funcție de diferite criterii, precum: categorii de vârstă, specializări profesionale, canale educaționale.
- Introducerea de programe de formare – atât din perspectiva funcționalității tehnologiei IA, cât și a securității cibernetice aferente – conform necesităților identificate în analiza menționată anterior.
- Identificarea programelor de formare existente și realizarea de demersuri pentru asimilarea acestora în circuitele formale de educație.
- Dezvoltarea și implementarea de programe de conștientizare (awareness) la nivelul decizionalilor din administrația publică.

- Realizarea unui ecosistem educațional (cu mecanisme corelate, de tip circuit auto-întreținut) care să permită educarea și antrenarea tinerilor în raport cu tehnologiile emergente, precum și inserția fluentă a acestora în piața muncii. De menționat că IA nu este prezentă doar ca specializare tehnică, ci modelează majoritatea domeniilor sociale și de viață.
- Dotarea instituțiilor de învățământ cu tehnologie care să permită formarea și exersarea competențelor legate de tehnologiile emergente.

Modalitățile prin care tehnologia IA poate poate crește performanța proceselor educaționale includ:

- IA pentru procesare informațională și managementul cunoașterii.
- IA pentru explorarea domeniilor științifice și realizarea de corelații inter-disciplinare.
- IA pentru îmbunătățirea conținutului și instrumentelor de practică (mai ales în corelație cu tehnologii precum Realitatea Virtuală și Realitatea Augmentată).
- IA pentru îmbunătățirea administrării proceselor educaționale (ex. IA care să asiste evoluția curriculum-ului educațional, IA care să asiste actualizarea setului de competențe umane dezirabile, IA pentru îmbunătățirea proceselor de evaluare, IA pentru facilitarea inserției absolvenților din învățământ în piața muncii).
- IA pentru îmbunătățirea managementului organizațiilor educaționale (ex. IA care să asiste profesorii în pregătirea și în derularea activităților educaționale, IA pentru îmbunătățirea administrării instituțiilor de învățământ, IA pentru organizarea proceselor educaționale).
- IA pentru identificarea și combaterea dezinformării.
- IA pentru asistență în luarea deciziilor.
- IA pentru stimularea creativității (care să asiste procesele de formare și management al talentelor).

Principalele reglementări de interes în această privință sunt reprezentate de legile care guvernează educația și cercetarea națională, precum și programele și mecanismele de finanțare a acestor domenii. Astfel, se identifică și necesitatea actualizării acestor documente legislative, spre a facilita implementarea reperelor mai sus menționate [82].

R Programele antiplagiat intră în grupa de risc ridicat deoarece afectează accesul la educație și evoluția profesională.

Inițiative relevante:

- ELLIS - European Labor for Learning and Intelligent Systems.
- În mai 2021 Parlamentul European a adoptat un Raport pentru utilizarea IA în educație, cultură și sectorul audiovizual [105]
- Digital Education Act New Curricula Framework for informatics (informatics for all)

Dezvoltarea resursei umane și pregătirea pentru transformarea pieței muncii. Sunt propuse inițiative politice [33], inclusiv:

- stabilirea de programe de educație formală în domeniul ingineriei și matematicii (STEM) și IA
- formare profesională și învățare continuă și programe legate de IA
- acordarea de sprijin financiar și nefinanciar pentru recalificarea și atragerea talentelor în IA
- încurajarea parteneriatelor academice între instituțiile de cercetare IA private și publice
- monitorizarea impactului IA pe piața muncii pentru intervenția politicilor.

4.10 Securitate cibernetică

La nivel internațional

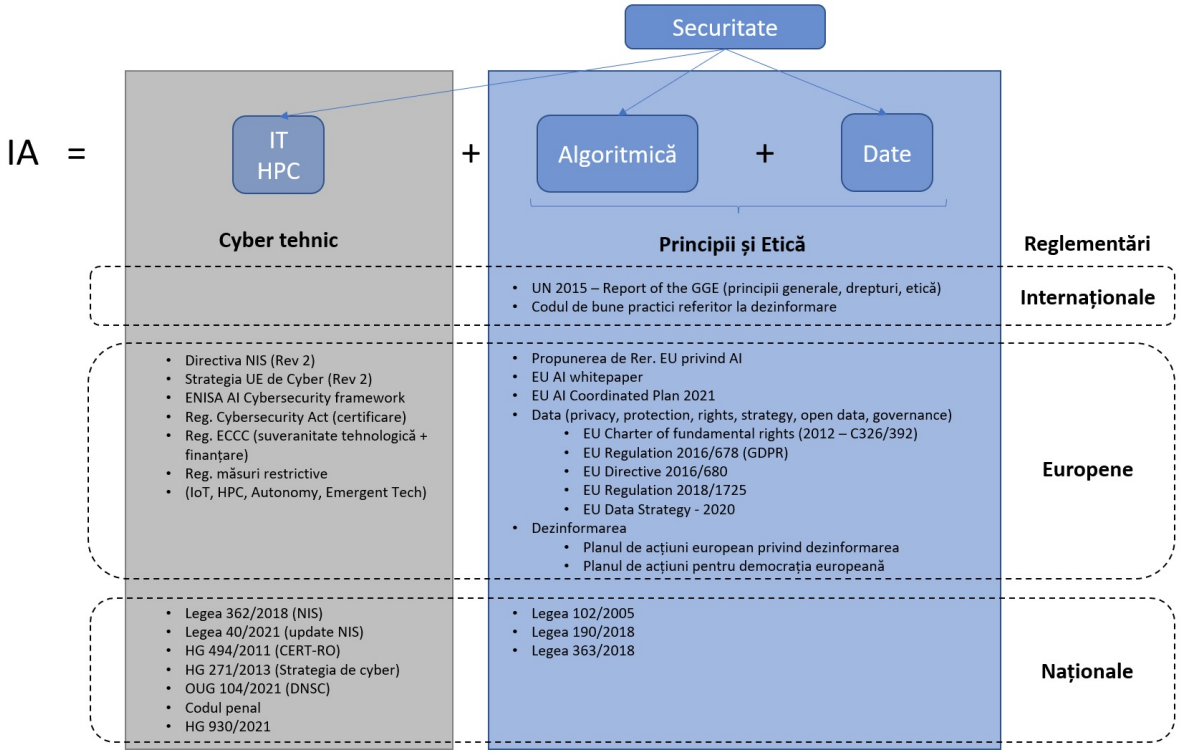

Figura 4.2: Securitate și IA etică

5 piloni ai rezoluției privind utilizarea tehnologiilor IT&C (71):

1. Amenințările existente și emergente
2. Dreptul internațional și libertățile fundamentale
3. Normele, regulile și principiile de utilizare a ICT. Aplicare voluntară pentru un comportament responsabil, la nivel național
4. Dezvoltarea încrederii
5. Cooperare internațională și asistență în dezvoltarea de capabilități

Norme internaționale pentru gestionarea ICTs: (71):

1. Cooperare inter-statală pentru menținerea stabilității și securității, și pentru prevenția practicilor distructive. Dezvoltarea și valorificarea de cadre/ instrumente de cooperare.
2. În caz de incidente: analiza de informații + analiza contextului de ansamblu + analiza efectelor generate. Evitarea interpretărilor greșite.
3. Responsabilitatea statelor de a interzice și de a preveni derularea de activități ilegale în mediul online.
4. Cooperare pentru tragerea la răspundere pentru activitățile ilegale.
5. Promovarea, protejarea și valorificarea drepturilor omului + dreptul la intimitate + libertatea de expresie
6. Interzicerea afectării infrastructurilor critice care oferă servicii publice.
7. Crearea unei culturi globale de securitate cibernetică și de protecție a infrastructurilor critice.
8. Statele trebuie să răspundă la solicitările de asistență.
9. Asigurarea integrității lanțului de aprovizionare. Prevenirea implementării de funcționalități ascunse.
10. Raportare responsabilă a vulnerabilităților și schimbul de informații privind soluțiile de remediere
11. Protejarea CERT-urilor altor state. Evitarea utilizării CERT-urilor pentru activități malițioase.

Sustenabilitatea presupune:
- Reglementări și politici
- Structuri și proceduri organizaționale
- Educație și instruire
- Echipamente (investiții și mentenanță) + resurse umane specializate
- Alocare de resurse financiare

R Codul de bune practici referitor la dezinformare – conține repere referitoare la prevenirea și combaterea dezinformării (e.g. fake news, deep fake)

Securitatea tehnologiei

Securitatea etapelor din ciclul de viață al tehnologiei (sau în întregul lanț de aprovizionare): cercetare, design, producție, certificare și conformitate, utilizare, recalibrare/ actualizare, analiza influențelor, scoatere din uz, reglementare, lanțul de aprovizionare.

Scuritatea categoriilor de componente ale produselor de IA: i) algoritmilor, ii) aplicațiilor/ software, iii) hardware (a puterii de calcul), iv) în cloud (a sistemelor si rețelelor de calcul), v) industrială și a infrastructurilor critice, vi) prevenția și reacția vii) controlul armelor viii) actualizarea legislației de securitate cibernetică, ix) actualizarea standardelor de securitate cibernetică.

Securitatea omului
- Siguranța, securitatea și sănătatea umane (infracțiuni de tip cyber-enabled și cyber-dependent)
- Protecția datelor
- Etica
- Dezinformarea (influențarea, amenințările hibride) [73]
- Protecția drepturilor omului
- Procesele de business și serviciile publice [și EU AI whitepaper, p.2]
- Analiza influențelor aduse omului (individual și colectiv)

Securitatea mediului
- Analiza influențelor aduse mediului/ naturii prin i) exploatarea resurselor, ii) utilizarea produselor finale, iii) generarea și gestionarea deșeurilor
- Principii și soluții pentru sprijinirea dezvoltării verzi

Guvernanța securității IA:
- Stabilite responsabilități concrete + sync sau mix cu mecanismele de cyber clasic
- Dezvoltarea de capabilități preventive de securitate IA (ex. detecție și urmărire)
- Dezvoltarea de mecanisme pentru certificarea (și monitorizarea) tehnologiilor IA
- Dezvoltarea de mecanisme pentru urmărirea evoluției riscurilor de securitate IA, schimbul de informații și coordonare la nivel național (civil-militar) și internațional
- Educație de securitate IA: conștientizarea populației cu privire la riscurile și beneficiile IA sau principii de securitate IA în sistemul educațional

La nivel UE

Aspecte principiale și de etică la nivelul UE:

Gândirea de securitate trebuie să sprijine obiectivele de bază ale tehnologiei IA, urmărind în principal asigurarea [EU AI whitepaper, p.9]:

1. Excelența – funcționalitate precisă și rate de erori cât mai mici
2. Încrederea – securitate ridicată și control asupra tehnologiei.

Complementaritate cu reglementările de securitate cibernetică și evitarea suprapunerilor.

1. Reglementările de cyber (dedicate) privesc aspectele tehnice (care sunt mai bazale și cu caracter de generalitate, în raport cu tehnologiile digitale).
2. Strategia națională IA și reglementările pe IA vizează chestiuni principiale de securitate.

Instrumente pentru guvernanța securității și conformității

- Cerințe legale
- Cod de conduită
- Principii de etică [extrase din EU AI whitepaper, 2020; EU AI Coordinated plan, 2021]: centrarea pe om, capacitatea umană de supraveghere a tehnologiei, responsabilitate și asumare, guvernanța datelor și a dreptului la intimitate, robustețe tehnică, siguranță și fiabilitate (cu valențe de securitate), transparență și explicabilitate, corectitudine și echitate. Diversitate, non-discriminare, calitate a vieții sociale și a mediului înconjurător.
- Indicatori/ criterii pentru evaluarea riscurilor
- Analiză de impact a riscurilor
- Digital trace (logging)

Planul coordonat UE privind IA, 2021:

- Facilitarea dezvoltării și a implementării.
- Sincronizarea laboratoarelor de cercetare cu mediul de producție comercial.
- Utilizarea IA în ca o forță a binelui: dezvoltarea talentelor, încrederea în IA, respectarea măsurilor tehnice prevăzute în [ENISA, 2020], respectarea prevederilor Strategiei UE de Securitate cibernetică [2020], IA sustenabilă și de încredere.
- Leadership strategic în sectoarele industriale cu impact de nivel înalt.

R Reglementări privind protecția datelor: e.g. 1 Carta drepturilor fundamentale a UE, GDPR, Directiva (UE) 2016/680, Regulamentul (UE) 2018/1725, Strategia europeană privind datele.

R Reglementări privind dezinformarea – centrate pe implementarea unor mecanisme eficiente de prevenire și combatere a creării, propagării și utilizării de informații false (fake news, deep fake, s.a.) Planul de acțiune european privind dezinformarea, Planul de acțiune pentru democrația europeană.

Aspecte tehnice

Directiva (UE) 2016/1148 privind măsuri pentru un nivel comun ridicat de securitate a rețelelor și a sistemelor informatice în Uniune (NIS), orientează eforturile pe o serie de sectoare industriale de interes principal și descrie reperele pentru dezvoltarea unei rețele piramidale de centre (europene și naționale) de tip CERT/ CSIRT/ SOC (Computer Emergency Response Team/ Computer Security Incident Response Team/ Security Operations Center). Varianta Rev.2 a Directivei NIS prevede extinderea ariilor/ domeniilor de implementare a măsurilor de securitate cibernetică și oferirea de responsabilități crescute autorităților competente.

EU Cybersecurity Strategy for the Digital Decade face refeire la trei instrumente; reglementări, investiții, politici, respectiv la trei arii de acțiune

1. Reziliența, suveranitate tehnologică și leadership
   - Directiva NIS
   - EU Cybershield – CSIRTs/ SOCs
   - Infrastructuri de comunicații ultra-securizate
   - Servicii de comunicații de bandă largă de tip next-gen
   - Un internet de lucruri securizate (certificate conform Regulamentului UE privind securitatea cibernetică)

- O securitate mai bună a internetului de scală globală
- Securitatea lanțului de aprovizionare tehnologic (Regulamentul UE privind ECCC)
- Forță de muncă specializată în securitatea cibernetică

2. Dezvoltarea capacității operaționale de a preveni și răspunde la incidente:
   - O unitate corelată de securitate cibernetică: autoritățile NIS + autoritățile de impunere a legii și judiciare + diplomație cibernetică + apărare cibernetică
   - Gestionarea criminalității cibernetice.
   - Instrumentar privind diplomația UE în domeniul securității cibernetice
   - Capabilități de apărare cibernetică
3. Spațiu global și deschis
   - Standarde, norme, politici (+ comportament statal responsabil/ principiile UN)
   - Cooperarea cu comunitatea
   - Capacitatea de reziliență globală

ENISA AI Cybersecurity, 2020:

- Cyber for AI vs. AI for Cyber vs. Malicious AI - controlul armelor, automatizarea cyber, acuratețea deciziilor
- Protejarea datelor, modelelor, artefactelor, actorilor, proceselor, mediului, uneltelor
- Robustețe, încredere, siguranță, transparență, explicabilitate, liabilitate, protecția datelor

Regulamentul (UE) 2019/881 privind ENISA (Agenția Uniunii Europene pentru Securitate Cibernetică) și privind certificarea securității cibernetice pentru tehnologia informației și comunicațiilor și de abrogare a Regulamentului (UE) nr. 526/2013 (Regulamentul privind securitatea cibernetică/ Cybersecurity Act), impune măsuri pentru:

- consolidarea stării de securitate cibernetică și creșterea capabilităților specifice, la nivelul statelor membre și al mediului de afaceri.
- dezvoltarea cooperării și coordonării între statele membre și instituțiile, agențiile și organismele europene.
- dezvoltarea capabilităților Uniunii de a reacționa în mod integrat la crize de securitate cibernetică transfrontaliere.
- creșterea culturii de securitate cibernetică la nivelul cetățenilor și a mediului de afaceri.
- creșterea transparenței privind modalitatea de asigurare a securității cibernetice a produselor și serviciilor IT&C, în vederea consolidării nivelului de încredere în piața unică digitală și în inovarea digitală.
- crearea unui cadru unitar de evaluare și certificare în domeniul securității cibernetice, la nivelul statelor membre, care să fie valabil pentru toate sectoarele industriale.
- consolidarea autorității europene în domeniu, ENISA – European Union Agency for Network and Information Security.
- asigurarea coerenței cu Regulamentul (UE) 2016/679 privind protecția datelor (General Data Protection Regulation – GDPR).

Regulamentul (UE) 2021/887 de instituire a Centrului de competențe european industrial, tehnologic și de cercetare în materie de securitate cibernetică și a Rețelei de centre naționale de coordonare (ECCC), stabilește:

- înființarea Centrului de competențe, a Rețelei de centre și a Comunității de competențe în securitate cibernetice, la nivel european.
- cerințele privind dezvoltarea și retenția capabilităților industriale și tehnologice de securitate cibernetică, necesare pentru securizarea Pieței Digitale Unice.
- dezvoltarea unui ecosistem european de cercetare – inovare – dezvoltare – pregătire în securitate cibernetică.
- promovarea și dezvoltarea cooperării în domeniul securității cibernetice la nivel european.

Regulamentul (UE) 2019/796 privind măsuri restrictive împotriva atacurilor cibernetice care reprezintă o amenințare la adresa Uniunii sau a statelor sale membre, stabilește repere de sancționare și răspuns la atacuri cibernetice semnificative, care sunt orientate împotriva infrastructurilor critice, a serviciilor necesare pentru menținerea unor activități sociale și/sau economice esențiale, a funcțiilor critice ale statului, a protecției informațiilor clasificate și a echipelor guvernamentale de răspuns la situații de urgență.

În plus, găsim relevanță pentru subiectul IA și în reglementările europene care privesc digitalizarea (ex. propunerea de Regulament pentru servicii digitale – 2020, propunerea de Regulament pentru piața digitală – 2020), redresarea și reziliența economică (Regulamentul UE 2021/241), protecția infrastructurilor critice (ex. propunerea de Directivă UE privind reziliența entităților critice - 2020), în viitor așteptându-ne să fie abordate și securitatea tehnologiilor autonome, securitatea dispozitivelor conectate (Internet of Things), securitatea rețelelor 5G și securitatea altor tehnologii emergente.

La nivel național

Găsim corespondență pentru reglementările europene privind datele, astfel: Legea 102/2005 privind înființarea, organizarea și funcționarea Autorității Naționale de Supraveghere a Prelucrării Datelor cu Caracter Personal, Legea 190/2018 de punere în aplicare a Regulamentului UE GDPR, Legea 363/2018 privind protecția persoanelor fizice referitor la prelucrarea datelor cu caracter personal de către autoritățile competente.

Legea nr. 362/2018 privind asigurarea unui nivel comun ridicat de securitate a rețelelor și sistemelor informatice, setează un cadru comun de cerințe de bază pentru securitatea sistemelor și rețelelor informatice ale operatorilor de servicii esențiale, respectiv furnizorilor de servicii digitale, la nivelul Uniunii Europene. Prin Legea 362/2018 sunt stabilite:

- autoritatea națională în domeniul securității sistemelor și rețelelor informatice, respectiv responsabilitățile acesteia.
- rolul autorităților naționale din diverse sectoare industriale și reperele de colaborare a acestora cu autoritatea națională în domeniul securității sistemelor și rețelelor informatice, precum și cu alte entități relevante la nivel internațional.
- cerințele de bază pentru securitatea sistemelor și rețelelor informatice, precum și managementul acestora.
- cadrul desfășurării activităților de audit în domeniu.
- promovarea culturii în domeniu, la nivel național.
- măsuri de supraveghere, control și sancționare privind respectarea legii.

Legea 40/2021 privind aprobarea Ordonanței de urgență a Guvernului nr. 119/2020 pentru modificarea și completarea Legii nr. 362/2018 privind asigurarea unui nivel comun ridicat de securitate a rețelelor și sistemelor informatice oferă completări și pârghii de impunere a Legii NIS (362/2018).

Hotărârea de Guvern nr. 271/2013 pentru aprobarea Strategiei de securitate cibernetică a României și a Planului de acțiune la nivel național privind implementarea Sistemului național de securitate cibernetică, stabilește repere de dezvoltare și coordonare la nivel național, pentru asigurarea securității cibernetice a infrastructurilor care sunt considerate critice pentru securitatea națională, buna guvernare, respectiv pentru maximizarea beneficiilor cetățenilor, mediului de afaceri și ale societății românești, în ansamblul ei. Concret, aceasta stabilește o serie de obiective și un plan de acțiune, vizând:

- completarea cadrului conceptual, de reglementare, organizatoric și acțional, necesar asigurării securității cibernetice la nivel național.
- dezvoltarea capacităților naționale de management al riscului și de reacție la incidente de securitate cibernetică.

- promovarea și consolidarea culturii de securitate în domeniul cibernetic.
- dezvoltarea unui Sistem național de securitate cibernetică (SNSC).
- respectiv reperele de cooperare între sectorul public și cel privat.

Hotărârea de Guvern nr. 494/2011 privind înființarea Centrului Național de Răspuns la Incidente de Securitate Cibernetică – CERT-RO descrie elemente tehnice de securitate cibernetică națională, precum și caracteristici ale structurilor de tip CERT.

Ordonanța de Urgență nr. 104/2021 pentru înființarea Directoratului Național de Securitate Cibernetică (DNSC) reformează domeniul securității cibernetice în plan național și înlocuiește CERT-RO cu DNSC, oferindu-i un registru mai extins de capabilități și responsabilități în specialitate.

Codul penal definește și incriminează anumite infracțiuni în domeniul securității cibernetice.

Hotărârea de Guvern nr. 930/2021 privind aprobarea Strategiei naționale împotriva criminalității organizate 2021-2024 include domeniul securității cibernetice printre ariile de interes care necesită acțiune strategică la nivel național.

4.11 Securitate și apărare națională

Tehnologia IA poate fi utilizată pentru a aborda problemele complexe de securitate și apărare cu care se confruntă România într-un context geo-politic, economic și social aflat în continuă evoluție. Terorismul, spionajul, frauda financiară la scară globală, știrile false și războiul informațional, amenințările cibernetice, amenințările hibrid și conflictele asimetrice sunt unele dintre provocările de nivel înalt în care ar putea fi folosite tehnologiile IA.

În același timp, tehnologiile digitale și emergente (inclusiv IA) prezintă riscul de a fi utilizate în scopuri dăunătoare și de a fi exploatate împotriva intereselor, sănătății, siguranței și securității societății și indivizilor. Soluțiile adaptate și inteligente pot valorifica tehnologiile IA pentru a contracara producerea de daune și pentru a împiedica proliferarea resurselor dăunătoare, pentru a asigura securitatea, siguranța, prosperitatea și bunăstarea populației.

Trei aspecte de bază care solicită atenție în ceea ce privește IA în relație cu securitatea și apărarea națională [12]:

IA pentru securitate: utilizarea tehnologiilor IA în scopuri de securitate și apărare națională (IA aplicată pe conținut de intelligence, pe date și informații care pot furniza contribuții utile pentru domeniul securității și apărării naționale). În acest sens, sursele care generează informații și conținut de lucru pentru securitatea și apărarea (comunitară și națională) sunt legate de dezvoltarea domeniilor sociale, de tendințele de evoluție tehnologică și de strategiile de abordare a piețelor comerciale. Astfel, menționăm ca relevante reglementările (europene și naționale, în vigoare și în curs de dezvoltare) care gestionează: digitalizarea, tehnologiile emergente, datele deschise, piața economică digitală, macro-finanțările pentru dezvoltarea tehnologiei s.a.

Securitate contra IA dăunător: mecanisme pentru prevenirea și contracararea utilizării tehnologiilor IA în scop dăunător. De interes, în mod special, sunt demersurile cu impact mare, strategice, care pot afecta securitatea și apărarea națională (ex. amenințări statale, rețele organizate, terorism, amenințări la adresa infrastructurilor critice). În această categorie intră:

- Reglementările de securitate și apărare națională (pentru România: HG 271/2013 privind Strategia națională de securitate cibernetică și propunerea de Lege a securității și apărării cibernetice)
- Reglementările de incriminare a acțiunilor ilicite (Regulamentul EU 2019/796, Codul Penal).
- Strategiile axate pe arii ale securității și apărării cibenretice, precum Strategia națională de apărare 2020-2024.

IA ca armă: utilizarea tehnologiilor IA în construcția armelor fizice și cibernetice este un domeniu sensibil, care pune în pericol securitatea și apărarea atât la nivel național, cât și comunitar și global, fiind gestionată prin instrumente și metode diplomatice în baza convențiilor internaționale și inter-statele, precum Raportul UN GGE 2015 privind tehnologia IT&C în contextul securității internaționale. Majoritatea formatelor de cooperare și diplomație abordează tematici privind controlul armelor cibernetice, mecanisme pentru dezvoltarea responsabilizarea statelor, transparența în comunicare și negocierea permanentă pentru evitarea conflictelor.

5. Protejarea împotriva dezinformării

Rezumat. Dezinformarea în România reprezintă o vulnerabilitate în contextul existenței unei populații cu educație eteorgenă și în contextul ignoranței inerente cu privire la tehnologii și informații tot mai complexe. Dezinformarea afectează direct funcționarea societății la scară largă, cu riscul de a afecta grav sănătatea sau democrația, așa cum o dovedește exemplul social recent de reticență la vaccinarea anti-Covid. Dezvoltarea prioritară de unelte de inteligență artificială pentru combaterea dezinformarii sau pentru educarea populației și micșorarea ignoranței poate aduce beneficii semnificative la nivelul societății.

Cadrul de reglementare pentru combaterea dezinformării rămâne o provocare datorită dificultăților de a trage linie clară între campanii de dezinformare și libertatea de expresie. Prezentăm aici inițiativele de reglementare și măsuri specifice referitoare la dezinformare În state precum Marea Britanie, Germania, Singapore, Cehia, Bulgaria, Moldova, Suedia, Rusia. De asemenea, abordarea industriei (i.e. platformele sociale) pe linia dezinformării este exemplificată prin prezentarea standardelor Facebook, a politicilor Google și Youtube sau a abordării TikTok. Sunt discutate aspecte legate de reglementarea generării de conținut sintetic video sau audio (i.e. deepfake), atât din pespectiva Comisiei Europene, cât și din cea a inițiativelor Facebook sau Google.

Pe de o parte, inteligența artificială este utilizată în acest context, pentru a crea conținut sintetic sau pentru a orchestra propagarea pe rețele sociale prin utilizatori roboți. Pe de altă parte, unelte IA au fost dezvoltate pentru detecția automată a conținutului fals și pentru amplificarea capabilitățile agentului uman de a analiza și verifica informații (e.g. fact checkers). La baza acestor unelte se află tehnologii precum procesarea limbajului natural, sumarizare automată, analiză semantică, analiză sentiment, extragere de entități sau algoritmi de identificare a conținutului sintetic. Pe această linie, prezentăm o serie de unelte IA folosite pentru combaterea dezinformării.

Capitolul se încheie cu o schiță de recomandări pentru combaterea dezinformarii în România prin înființarea unui birou specializat în cadrul Consililului Național al Audiovizualului, în colaborare cu Autoritatea de Reglementare a Inteligenței Artificiale.

Fenomenul dezinformării, cunoscut în mod colocvial ca “fake news” (i.e., știri fabricate), este definit de către unitatea Media Convergence and Social Media al Comisiei Europene ca fiind informații false sau înșelătoare verificate create, prezentate și diseminate în scopul câștigului economic sau pentru a înșela publicul în mod intenționat [64]. Acest fenomen este reconoscut ca având un impact semnificativ la nivel global asupra stării democrației și exercițiului electoral, a polarizării artificiale a dezbaterilor, și cu riscul de a afecta grav sănătatea, mediul și siguranța

cetățenilor Europeni.

R În contextul pătrunderii inteligenței artificiale în întreaga societate este posibilă apariția unei mișcări anti-IA care să fie susținută prin campanii de dezinformare care utilizează unelte IA.

5.1 Reglementări și inițiative în UE

Legislația și practicile curente din domeniul audiovizualului au competență și aplicabilitate limitată în cazul mediilor online. La nivelul Uniunii Europene există însă diferite inițiative menite să limiteze impactul dezinformării asupra populației uniunii, mai ales în mediul online:

Raportul Comisiei Europene pe linia dezinformării [14] recomandă cinci direcții de acțiune:

1. asigurarea transparenței prin partajarea de informații cu privire la sistemele care asigură circulația online a știrilor
2. educarea consumatorilor de știri online
3. dezvoltarea de unelte pentru utilizatori și ziariști pentru a identifica dezinformarea
4. sustinerea diversității și sustenabilității ecosistemului EU de știri.
5. cercetarea continuă a impactului dezinformării

Aceste acțiuni sunt suplimentate în comunicare Comisiei Europene cu privire la Tackling Online Disinformation de:

1. Acțiuni ale platformelor online (e.g. facilitarea evaluării de către utilizator a conținutului
2. Rebalansarea relației între media și platformele online
3. Cooperarea dintre verificatorii de conținut (i.e. fact-checkers) independenți
4. Aplicarea de tehnologii bazate pe IA pentru a combate dezinformarea
5. Susținerea jurnalismului de calitate cu ajutoare de la statele membre

Network Enforcement Act (i.e. NetzDG) împotriva discursului urii (Introduce legislație specifică impotriva discursului urii cu efect din Ianuarie 2018 (e.g. termene clare pentru eliminarea conținutului, cuantumul amenzilor) [47], [89]

- Codul de bune practici privind dezinformarea stabilește un set de standarde de auto-reglementare la nivel mondial pentru industrie (Code of Practice on Disinformation | Shaping Europe's digital future;
- Observatorul european pentru mass-media digitală este un hub european pentru verificatori de date, cadre universitare și alte părți interesate relevante pentru a sprijini factorii de decizie politică (EDMO – United against disinformation);
- planul de acțiune privind dezinformarea vizează consolidarea capacității și cooperării UE în lupta împotriva dezinformării (Action Plan against Disinformation - European External Action Service (europa.eu));
- Planul de acțiune pentru democrația europeană va elabora orientări pentru obligațiile și responsabilitatea platformelor online în lupta împotriva dezinformării (European Democracy Action Plan | European Commission (europa.eu);
- Comunicarea privind „abordarea dezinformării online: o abordare europeană" este o colecție de instrumente pentru a combate răspândirea dezinformării și pentru a asigura protecția valorilor UE (Communication - Tackling online disinformation: a European approach | Shaping Europe's digital future (europa.eu));
- Programul de monitorizare și raportare COVID-19, desfășurat de semnatarii Codului de bune practici, acționează ca o măsură de transparență pentru a asigura răspunderea în abordarea dezinformării (First baseline reports - Fighting COVID-19 disinformation Monitoring Programme | Shaping Europe's digital future (europa.eu)). Semnatari ai Codului de Bune practici au pus la dispoziția Comisiei Europene rapoarte legate de activitățile intreprinse

pentru a stopa dezinformarea legată de vaccinuri. Facebook, Google, Microsoft, Twitter, TikTok si Mozilla au luat o serie de masuri, tehnologice dar și de verificare factuală (i.e., fact checking) cu profesioniști (Reports on January actions – Fighting COVID-19 Disinformation Monitoring Programme | Shaping Europe's digital future (europa.eu)). Măsuri relevante pentru România au luat doar Microsoft si Google pentru motoarele lor de căutare, Bing și Google Search, implementând microsite-uri de tip information hubs cu știri veridice.

5.2 Politici publice împotriva dezinformării

Pachete de legi și măsuri specifice referitoare la dezinformare au fost implementate în unele state naționale, atât în Europa cât și la nivel global.

Germania a adoptat în iunie 2017 Network Enforcement Act, prin care companiile media și rețelele sociale care nu șterg conținutul ilegal (asimilat și conținut fals identificat ca atare de autorități) în 24 ore de la primirea unei plângeri și blocarea conținutului ofensator în 7 zile de la notificare, cu amenzi de până la 5 milioane de euro.

Mare Britanie a înființat Fake News Rapid Response Unit (RRU) în iulie 2018, ca parte a guvernului (Cabinet Office) ce lucrează în colaborare cu National Security Communications Team (NSCT) pentru a identifica conținut fals în mediul online, în principal pe rețelele sociale[a]. Unitatea se ocupă în prezent și cu combaterea dezinformării legate de pandemia de coronavirus, cu peste 70 de cazuri săptamânale în 2020 (Government cracks down on spread of false coronavirus information online - GOV.UK)

[a] https://www.pressgazette.co.uk/prime-minister-announces-rapid-response-unit-to-tackle-fake-news

Singapore a extins în 2019 responsabilitățile Infocomm Media Development Authority (IMDA), parte a Ministerului Comunicațiilor și Informației, organizație similară cu Consiliul Național al Audiovizualului (CNA), prin POFMA (Protection from Online Falsehood and Manipulation Act). Modelul implementat este unul de licențiere a site-urilor web similar cu cel al posturilor TV, acordând licențe de funcționare website-urilor care: (a) publică cel puțin un articol pe săptămână despre Singapore pe o perioada de 2 luni și (b) au 50,000 de IP-uri din Singapore trafic pe luna, pentru o perioadă de cel puțin 2 luni. POFMA [84] are rol și de a defini bune practici, pregatire și colaborare cu mediul academic și de business în abordarea dezinformării online.

Republica Cehă participă în East StratCom Task Force iar din 2018 este parte a NATO Cooperative Cyber Defence Centre of Excellence. Servicii de informații (BIS și National Cyber and Information Security Agency (NUKIB)) emit avertismente clare de fiecare dată când o compania de dezinformare se derulează. Din punct de vedere instituțional, monitorizarea dezinformării este realizată de Centre Against Terrorism and Hybrid Threats. Hanzelka et al. notează că strategia națională pentru combaterea dezinformării include educația la nivel universitar (atât cel militar, cât și civil), dar nu este vizibilă o cooperare cu media [45].

Bulgaria are campanii de dezinformare care reiterează mesaje de tipul "EU as a US-NATO institution", sau "EU decline under immigration". Riscurile asociate dezinformării nu sunt menționate de către State Agency for National Security, iar măsuri legale specifice nu există [45].

România nu are nicio autoritate sau departament de stat dedicat exclusiv luptei împotriva dezinformării. Consiliul Național al Audiovizualului (CNA) este în acest moment singura autoritate care poate lua măsuri directe împotriva dezinformării din mediile TV și Radio, în urma unor sesizări, potrivit Legii

audiovizualului Nr. 504 din 11 iulie 2002 și a Codului de reglementare a conținutului audiovizual. Autoritatea CNA se răsfrânge asupra posturilor radio și TV cu licență audiovizuală a căror emisii sunt monitorizate și analizate cu ajutorul unor agenții subcontractate (Kantar Media, ARMA), rezultatul încălcării legii și a regulamentului fiind amenzi, scoaterea de pe post a unor secvențe audio/video, sau retragerea licenței de emisie. CNA publică în mod transparent deciziile luate de membrii Consiliului, decizii supuse la vot, care pot fi apoi contestate în instanță. CNA nu acoperă prin jurisdicție mediul online, unde fenomenul dezinformării e cel mai răspândit.

Serviciul Român de Informații (SRI), chiar dacă nu are autoritate de control al mediului online, întreprinde campanii de conștientizare a fenomenului ”fake news“, cum ar fi programul AWARENESS.

R În România, campanii de dezinformare reiterează, printre altele, mesaje de tipul: ”întoarcerea la valorile tradiționale creștine“, atacul asupra celor care impun ”valori străine“, ”țară de mâna a doua, sau “stadardele duble ale UE” [62]. Conley et al. prezintă corupția și religia ca două puncte de intrare pentru campaniile de dezinformare din România [31].

Moldova. Codul Audiovizualului din Moldova adoptat în 2018, sau ”legea anti-propagandă“ cum a fost denumit, pare să nu aibă niciun impact, deoarece campaniile de propagandă s-au mutat de la știri în programele de divertisment. Lecția pentru următoarele politici este că focalizarea doar pe conținut nu este suficientă, fiind necesară evaluarea generală a furnizorului de media[a].

[a] https://euvsdisinfo.eu/moldova-parliament-elections-how-disinformation-did-not-work

Suedia se opune dezinformării – denumită în strategia națională ca ”information influence“ – prin Swedish Civil Contingencies Agency (MSB). Cum strategia de apărare are în centru conceptul de ”reziliență“, în realitate agenții guvernamentale diferite combat dezinformarea fără implicarea guvernului. Un dezavantaj al acestei *abordări bottom-up* este nevoia de coordonare. Conceptul de ”reziliență” este implementat pe mai multe paliere: 1) la nivelul populației, 2) la nivelul mediului de afaceri, respectiv 3) în administrația publică.

În primul rând, la nivelul societății, reziliența este instanțiată prin educarea populației. De exemplu, începând cu 2018, elevii din ciclul primar sunt învățați cum sa distingă între surse de informare de încredere de cele nesigure [106]. Educarea populației pe această linie este responsabilitatea Swedish Media Council. Scopul acestor politici de educare extensivă corelate cu acțiuni de managmentul crizelor urmăresc contruirea unui bariere psihologice. Din 2022 apărarea împotriva dezinformărilor va fi coordonată de Agency for Psychological Defence, desinformarea fiind privite ca o dimeniune dintr-un concept mai larg “psychological warfare” [103]. Totuși, responsabilitatea de a implementa strategia națională rămâne la nivelul municipalităților, păstrând linia de apărare decentralizată a strategiei pentru reziliență.

În al doilea rând, pe lângă educarea populației, autoritățile suedeze se focusează și pe educarea în companii, deoarece firmele au fost identificate ca vulnerabilități și puncte de intrare pentru campaniile de dezinformare.

În al treilea rând, pentru administrația publică s-a pus la dispoziție un manual (Countering Information Influence Activities: A Handbook for Communicators [76]) pentru identificarea și combaterea știrilor fabricate. Manualul include descrieri ale instrumentației tehnice pentru manipulare precum agenți software (boți), deepfakes, învățare automată, phishing, sockpuppets [77]. De exemplu, funcționarii publici sunt ghidați cum să identifice un bot pe baza a șapte trăsături: i) poza de profil (verificarea autenticității prin căutare de imagini); ii) activitatea (sunt enumerate euristici precum “boții sunt extrem de activi, cu mai mult de 50 de mesaje pe zi”); iii) numele (numele generate automat pot conține litere aleatoare); iv) data de creare a contului (boții nu au istorie, iar perioadele de inactivitate totală alternează cu cele de activitate intensă); v) limbaj (boții utilizează traduceri automate sau pot posta același conținut în diferite limbi); vi) informații (boții pot utiliza informație fictivă sau furată) vii) implicare (boții răspândesc de multe ori informații preluate de la alți boți). Pe lângă aceste indicii de natură tehnică, manualul exemplifică strategii retorice și persuasive utilizate de troli sau “shills”: ad hominem, whataboutism, gish-gallop, strawman, hijacking. În plus, manualul prezintă tehnici avansate de coordonare precum polarizarea, spălare de bani,

provocări sau flooding [77]. După clarificarea acestor trei dimensiuni – manipulari tehnice, elemente de retorică, tehnici de coordonare – sunt furnizate tehnici de combatere a dezinformării. Sunt prezentate moduri de a formula *răspunsuri bazate pe fapte* sau *răspunsuri bazate pe argumente*. În final sunt prezentate modalități prin care funcționarii publici să partajeze și sa propage cazurile cu care s-au întalnit. Această partajare a exemplelor de bune practici este un element important în strategia decentralizată, bottom-up a Suediei.

MSB încadrează activitățile de cercetare specifice ca o contramasură operativa impotriva dezinformării. Astfel, Swedish Defence Research Agency (FOI) are ca sarcină și monitorizarea activităților automate în media online (i.e făcute de boți sau altă instrumentație bazată pe IA). Pe aceeasi direcție, Lund University are sarcina de a concepe ghiduri și cursuri pentru a combate dezinformarea [106].

R Atât Suedia cât și Singapore au adoptat abordarea "Apărare totală".

R Vilmer et al. argumentează că unele practici pentru combaterea dezinformării depind de context: ce funcționează în Suedia datorită încrederii ridicate a cetațenilor în autorități s-ar putea să nu funcționeze în societăți puternic polarizate precum cea din SUA [106].

Rusia privește dezinformarea ca o instanță a "confruntării informaționale". Pentru Russia Armed Forces, *războiul psihologic* reprezinta "influențarea psihologică și informațională a audienței din străinătate cu scopul de crea opinii și comportamente în interesul național al Rusiei" [95].

Piața neagră a manipulărilor prin media online este în expansiune odată cu creșterea numărului de furnizori și cumpărători de care se întâlnesc să tranzacționeze clickuri, aprecieri, commentarii, urmăritori [99]. Se pot cumpăra de exemplu unelete și servicii petnru manipularea media care includ și servicii de suport pentru aspecte precum

- Creare conturi false: prețul depinde de modul în care contul a fost creat (automat, manual sau furat de la utilizatori), de calitatea conținutului de pe acel cont, vechimea contului (care poate varia de la câteva zile până la 7 ani de exemplu)
- Manipularea metricilor sociale: prin utilizarea de conturi false, platforme cu persoane care lucrează pe mai puțin de 1$ pe oră, schimburi de apreciere pozitive, utilizare software malițios
- DDoS2.0: noile atacuri distribuite de tipul Denial-of-Service attacks în cadrul paginilor de pe platformele media, țintele obișnuite fiind activiști sau jurnaliști. Prețurile pentru astfel de servicii DDoS pornesc de la $5 și ajung la $200 [99].

EUvsDiInfo este proiectul fanion al European External Action Service's East StratCom Task Force. EUvsDiInfo furnizează știri și analize (disponibile inclusiv în limba rusă) legate de știri fabricate. Este furnizată baza de date Disinfo cu 12,472 cazuri de dezinformare care pot fi filtrate în funcție de țară, dată sau subiect (e.g. 919 cazuri privesc dezinformarea legată de COVID-19). Pentru a fi inclus în baza de date un mesaj trebuie să fie îndeplinite simultan două criterii: (i) să fie verificat ca fals sau manipulator pe baza unor dovezi factuale disponibile public, și (ii) să fie emis de o sursă media ale cărei fonduri provin de la Kremlin. A compilație săptămânală colectează principalele cazuri de dezinformare pro-Kremlin, regulat fiind conceptute studii educative împotriva dezinformării. Pe această linie educațională, sunt furnizate jocuri în care utilizatorii se pot antrena pentru a identifica știrile fabricate. Similar, EU DisInfo Lab furnizează o serie de unelte web pentru monitorizarea dezinformării.

5.3 Abordarea dezinformării de către Big Tech

Oamenii interacționează în mod constant cu inteligența artificială, direct și indirect, chiar și fără ca aceștia să conștientizeze acest lucru: fie că utilizează IA prin intermediul dispozitivelor inteligente

Tabela 5.1: Activități pentru combaterea dezinformării [45]

Categoria	Activități
Acțiuni cu public	Atenționări clare (i.e. "clear warnings" emise de serviciile de informații , comunicate publice (actorii politici), educarea votanților (e.g. Swedish Civil Contingencies Agency 2018), educarea părților
Măsuri la nivelul statelor	Instituții la nivel ministerial, educarea angajaților
Măsuri legale	implementarea măsurilor legale, cooperare cu mediul economic
Acțiuni cu media	Antrenarea comunicatorilor, monitorizarea dezinformării, susținerea jurnalismului de investigației
Acțiuni cu partidele politice	Atragerea partidelor în cultura politica a respectării normelor, educarea persoanelor implicate in campaniile electorale
Contramăsuri directe	Operații informaționale împotriva actorilor ostili, presiune diplomatica
Acțiuni cu entitățile supranaționale	Forțe comune de cooperare, discuții comune, acceptarea recomandărilor

de uz casnic - televizoare inteligente, frigidere, aspiratoare robot, dispozitive inteligente controlate prin voce și dispozitive mobile, pentru a interacționa cu furnizorii de servicii publice sau utilități, prin intermediul instrumentelor pentru petrecerea timpului liber, învățare sau muncă, cum ar fi soluții de securitate a rețelelor, sisteme de conferințe web, software eLearning și recrutare, și monitorizare a performanței în domeniul resurselor umane.

Cel mai probabil, IA are cea mai mare impact asupra cetățenilor atunci când aceștia interacționează cu software "Big Tech", cum ar fi Google, Facebook, Apple, Microsoft, Tiktok și Amazon. În ceea ce privește gestionarea informațiilor online, Google este cel care deține 86.6% din piața de căutare pe Internet (Internet Search) începând cu 2021 februarie și cu 2.3 miliarde de utilizatori pe Youtube, reprezentând 79% din utilizatorii de Internet (https://www.oberlo.com/blog/youtube-statistics), doar Facebook având utilizatori mai activi (aproape 2.8 miliarde). Potrivit Digital News Report din 2021 publicat de Institutul Reuters al Oxford University, Facebook este folosit săptămânal pentru știri de către 44% utilizatori, urmat de Youtube (29%) și Twitter (13%), ceea ce poate duce la concluzia că Google și Facebook sunt principalele platforme prin care informațiile sunt distribuite și consumate la nivel global. Impactul standardelor, protocoalelor și instrumentelor acestor platforme de combatere a dezinformării este semnificativ și nu ne putem imagina că lupta eficientă împotriva știrilor false se poate face fără implicarea acestor actori.

Guess et al. argumentează că "Facebook este de departe cel mai rău făptuitor atunci când vine vorba de răspândirea de știri false. Mai rău decât Google. Mai rău decât Twitter. Și mai rău decât furnizorii de webmail, cum ar fi AOL, Yahoo!, și Gmail." Studiul a urmărit 3000 de americani înaintea alegerilor prezidențiale din 2016 din SUA și a găsit Facebook ca referent pentru surse de știri de neîncredere în peste 15% din cazuri. De la alegerile prezidențiale din 2016 din SUA, Facebook a pus în aplicare mai multe proceduri, instrumente și inițiative de combatere a știrilor fabricate [44]

Standardele Facebook

Facebook a devenit principalul actor datorită poziției sale dominante în canalele de distribuție a știrilor pe bază de algoritmi [64].

Termenii de utilizare Facebook (`https://www.facebook.com/terms.php`) precizează în

mod clar serviciile oferite prin intermediul platformei, pe lângă scopul inițial de „conectare cu oamenii" (connecting with people): Crearea de conținut („exprimați-vă și comunicați cu privire la ceea ce contează pentru dvs.") și consumarea de conținut („descoperă conținut") gratuit, în timp ce este deservită publicitate care constituie principalul flux de venituri pentru Facebook. Aceste servicii sunt guvernate de o politică de „combatere a conduitei dăunătoare și protecție și sprijinire a comunității noastre" printr-o combinație de echipe dedicate și sisteme tehnice care pot duce la eliminarea conținutului.

Standardele comunitare oferă detalii suplimentare cu privire la activitățile permise pe Facebook, care vizează domenii precum Violența și Comportamentul Penal, Siguranța, Conținutul Inacceptabil, Integritatea și Autenticitatea, Protecția Proprietății Intelectuale, și Cererile și Deciziile legate de Conținut. Secțiunea Integritate și Autenticitate se bazează pe ceea ce se numește „[autenticitate este] piatra de temelie a comunității noastre" și valorifică „o combinație de sisteme automate și manuale pentru a bloca și elimina conturile care sunt utilizate pentru a abuza în mod persistent sau flagrant de standardele noastre comunitare". Infracțiunile repetate împotriva standardelor comunitare ale Facebook pot duce la eliminarea conturilor, care sunt precedate de „o șansă de a învăța regulile noastre și de a urma standardele noastre comunitare".

În domeniul specific al știrilor fabricate, politica Facebook (https://www.facebook.com/communitystandards/false_news) este de a reduce distribuția, mai degrabă decât eliminarea, pentru că "recunoaștem, de asemenea, că aceasta este o problemă dificilă și sensibilă. Vrem să ajutăm oamenii să rămână informați fără a sufoca discursul public productiv. Există, de asemenea, o linie fină între știri false și satiră sau opinie".

Strategia Facebook pentru oprirea știrilor fabricate datează din 2018 și are trei părți:

1. Eliminarea conturilor și conținutului care încalcă standardele comunitare sau politicile de publicitate: i) se verifică identitatea conturilor și activitatea acestora, ii) asigurarea transparenței sursei unei reclame politice, sau iii) eliminarea boților;
2. Reducerea distribuirii de știri fabricate și conținut inautentic, cum ar fi Clickbait, limitarea distribuției la prieteni-de-gradul-întâi[a], introducerea unui mecanism de control a calității știrilor bazat pe modele de încredere [64].
3. Informarea oamenilor, oferindu-le mai mult context cu privire la postările pe care le văd

[a] Deși acest lucru poate amplifica efectul "camerelor- de ecou"

R În 2018, Facebook estima că platforma este infestată de 60 milioane bots [58].

Se menționează că "știrile false nu încalcă standardele noastre comunitare", dar dacă distribuirea de informații false este corelată cu încălcarea altor reguli (cum ar fi intimidarea, imitarea, reclamele frauduloase) sau este considerată spam, se vor lua măsuri împotriva acestor conturi, inclusiv a postărilor conturilor respective. Comportamentul neautentic este detectat și prin utilizarea învățării automate (i.e. ML): „de asemenea, folosim învățarea automată pentru a ajuta echipele noastre să detecteze frauda și să pună în aplicare politicile noastre împotriva spam-ului. Acum blocăm milioane de conturi false în fiecare zi când încearcă să se înregistreze".

Informații despre alte eforturi Facebook în lupta lor împotriva dezinformării sunt publicate pe pagina about.fb.com , cu subiecte specifice, cum ar fi alegerile generale naționale, luarea de măsuri împotriva persoanelor care distribuie în mod repetat informații eronate, vaccinuri și Covid19, schimbări climatice și aplică și limitări de funcționalitate în aplicația Messenger.

Fact-checking și tehnologii pentru detectarea informațiilor false Tactica de combatere a știrilor false ia în considerare, de asemenea, monetizarea anunțurilor pe pagini cu calitate scăzută (numite ferme de publicitate - ad farms), penalizarea clickbait, și link-uri partajate frecvent de spammeri. Adresarea

problemei titlurilor de tip Clickbait în Facebook News Feed nu este detaliată în niciunul din articolele legate de calitatea fluxului de știri (https://about.fb.com/news/2016/06/building-a-better-news-feed-for-you/, https://about.fb.com/news/2017/06/news-feed-fyi-showing-more-informative-links-in-news-feed/), dar filtrarea titlurilor care sunt clickbait ar trebui sa folosească un algoritm automat datorita volumului mare de date, deși acest lucru nu este menționat nicăieri în mod explicit („ne vom uita acum dacă un titlu reține informații sau dacă exagerează informații separat [...] vom continua să învățăm în timp, și sperăm să continuăm extinderea acestei lucrări pentru a reduce clickbait în mai multe limbi" (https://about.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/).

Folosirea de unelte externe de tip fact-checkers este un efort constant al Facebook în lupta împotriva știrilor false, efort combinat cu inițiativele educaționale și de alfabetizare. Fact-checker-ii sunt certificați prin intermediul organizației independente International Fact- Checking Network. Efortul unui fact-checker (verificator de fapte), care marchează o știre ca fiind falsă, va duce la deprioritizarea știrii în News Feed, reducând vizualizărileviitoare ale știrii respective cu 80%. Informațiile de la uneltele fact-checkers sunt, de asemenea, utilizate pentru a îmbunătăți tehnologia Facebook, probabil prin crearea de date utilizate în procesul de învățare computațională.

■ **Exemplu 5.1 — Centrul de știri al Facebook în context electoral.** Așa cum este descris în postarea Facebook legată de alegerile din Etiopia din 2021 (https://about.fb.com/news/2021/06/how-facebook-is-preparing-for-ethiopias-2021-general-election), Facebook operează un Centru de știri pentru alegeri, înființat în 2018, numit "Camera Războiului" (War Room) (https://about.fb.com/news/2018/10/war-room/) pentru a lupta împotriva interferențelor electorale în timp real. Echipa globală care lucrează în zona de siguranță și securitate a Facebook este de 35.000 de persoane („în ultimii ani, am triplat dimensiunea echipei globale care lucrează la siguranță și securitate pentru peste 35,000 și a angajat mai mulți referenți de conținut care sunt vorbitori nativi de Amharic, Oromo și Somali, având în același timp capacitatea de a revizui conținutul în limba Tigrinya.") care sunt active în eliminarea discursurilor de incitare la ură și a altor conținuturi și conturi care utilizează „tehnologia de detectare proactivă care identifică conținut de tip hate speech sau violență și incitare".

De asemenea, în contextul monitorizării alegerilor și al alegerilor etiopiene din 2021, în special, Facebook elimină în mod activ „cele mai grave tipuri de dezinformare, cum ar fi conținutul destinat să suprime votul sau care ar putea provoca violență sau daune fizice", în timp ce pentru alte tipuri de conținut se folosește de parteneriate cu organizații de fact-checking extern "pentru a stabili dacă ceva este dezinformare sau știri false. Atunci când analizăm și clasificăm o informație ca fiind falsă, reducem distribuția acestuia, astfel încât mai puțini oameni o mai văd și adăugăm o etichetă de avertizare cu mai multe informații pentru oricine o vede. În general, atunci când un ecran de avertizare este plasat la o știre, 95% din timp oamenii nu mai acesează respectiva informație. "

■

Nu este clar dacă aceste campanii de monitorizare a alegerilor organizate conduse de Facebook au sprijin din partea unor reprezentanți naționali oficiali sau dacă echipa de 35,000 de membri este implicată în monitorizarea campaniilor electorale din alte țări decât cele menționate pe site-ul lor. De asemenea, în ce măsură organizațiile independente de verificare a informațiilor sunt libere de orice legături politice sau legate de guvern. Un exemplu este AFP (Agence France Presse), o agenție de știri care operează o rețea de verificare a informațiilor pentru Facebook, care este activă în monitorizarea alegerilor, agenție care este controlată de guvernul francez, deși funcționează ca o companie independentă în Franța (www.referenceforbusiness.com/history2/57/Agence-France-Presse.html#:~:text=Agence%20France%2DPresse%20(AFP),of%20its%20top%20media%20clients).

Utilizarea IA în acest context este foarte clar menționată de Facebook, deși detaliile privind rezultatul și impactul algoritmului sunt greu de determinat la nivel individual: „lucrăm cu echipa noastră de cercetare IA, învățând din mediul academic, extinzând parteneriatele noastre cu fact-

checkeri externi și vorbind cu alte organizații - inclusiv cu alte platforme - despre cum putem lucra împreună”[1].

Detaliile tehnice despre învățarea automată utilizată pentru algoritmul de ierarhizare a fluxurilor de știri din News Feed sunt disponibile pe blogul de inginerie Facebook, dar fără nicio informație specifică despre modul în care informațiile false afectează clasamentul știrilor, cu excepția reducerii afișărilor informațiilor marcate ca fiind false, așa cum s-a menționat anterior. Algoritmii determină ce știri apar pentru fiecare din cele 2 miliarde de oameni ce folosesc Facebook News Feed, cu un feed unic pentru fiecare utilizator, folosind Machine Learning, cum ar fi învățarea tip multitasking pe rețele neuronale, embeddings, și sisteme de învățare offline.

Întrebări cu privire la acuratețea atât a algoritmilor de clasificare a fluxului de știri, algoritmii utilizați pentru a detecta automat comportamentul inautentic, clickbait, și știri false, și imparțialitatea fact-checkerilor independenți sunt importante pentru utilizatorii individuali care se bazează pe Facebook pentru a-și promova afacerile, cauzele politice sau sociale, dar și în ceea ce privește informațiile la care au acces. A fi aruncat în “închisoarea Facebook”, așa cum este colocvial menționată această situație pe Internet, poate însemna având un cont temporar sau permanent suspendat din cauza încălcării standardelor comunitare, uneori fără a obține detalii cu privire la infracțiunea exactă, sau având o modalitate clară de a contesta decizia Facebook. Facebook oferă opțiunea de a apela (https://www.facebook.com/help/346366453115924/) deciziile referitoare la conținutul propriu sau raportat unui Consiliu de Supraveghere (Oversight Board) (https://www.facebook.com/help/711867306096893?helpref=related&ref=related&source_cms_id=346366453115924). Rezultatul poate fi o răsturnare a unei decizii Facebook, dar, spre deosebire de instituțiile juridice din întreaga lume, care sunt mandatate să ia în considerare analiza oricărei revendicări, “nu toate deciziile de conținut sunt eligibile pentru apel la Consiliul de supraveghere”.

Detectarea Similarității Imaginilor

Facebook abordează, de asemenea, dezinformarea folosită în imagini folosind o soluție numită SimSearchNet++. Facebook explică modul în care a dezvoltat un sistem IA care poate detecta similarități (near duplications) ale imaginilor ce conțin știri false dovedite și se poate ocupa de manipulări precum redimensionări/decupări, blurări și capturi de ecran, inclusiv imagini care conțin text. Instrumentul este utilizat pentru a trimite avertismente utilizatorilor înainte de a interacționa cu conținut identificat deja ca fiind potențial fals, la scară mare, cu 180 milioane de imagini care primesc etichete de avertizare în SUA în perioada martie - iunie anterioară alegerilor din SUA din 2020, pornind de la conținutul identificat manual de fact-checkeri.

SimSearchNet este un model neural convoluțional (convolutional neural network CNN) construit special pentru a detecta near-exact duplicates, care a fost proiectat ca parte a efortului Facebook de indexare și comparare a imaginilor și poate detecta imaginile duplicate parțial. SimSearchNet este o tehnologie proprietară a Facebook și folosește computer vision "în special, pe construirea de reprezentări compacte care ne permit să indexăm și să căutăm rapid fotografii la scară [101]. Capacitatea de căutare a fotografiilor este disponibilă ca FAISS (Facebook AI Similarity Search)[a] pentru a fi utilizată de dezvoltatorii terți care necesită o bibliotecă pentru a căuta rapid embeddings în documente multimedia care sunt similare între ele.

[a] https://ai.facebook.com/tools/faiss

■ **Exemplu 5.2 — Utilizare SimSearchNet: Covid19.** SimSearchNet a fost folosit în mai 2020 pentru a identifica conținutul manipulativ referitor la Covid19 (https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content), vizând oprirea exploatării financiare a situației de urgență prin detectarea și eliminarea publicității false pentru produse precum măști, agenți de curățare și kituri de testare falsificate (Figura 5.1). Algoritmii utilizează o bază de date object-level care conține anunțuri semnalizate manual care

[1] https://about.fb.com/news/2018/05/hard-questions-false-news

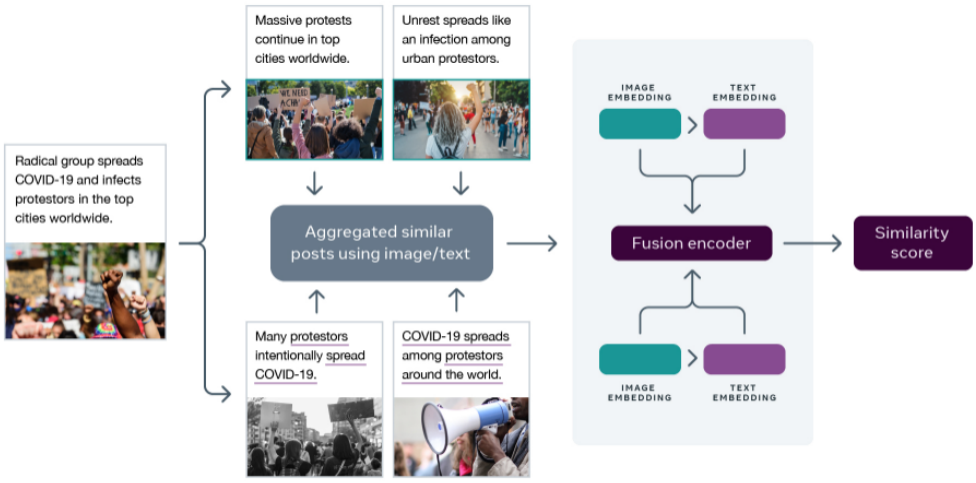

Figura 5.1: Utilizare SimSearchNet pentru a identifica conținutul manipulativ referitor la Covid19

încalcă politica Facebook. Această bază de date este apoi utilizată pentru a detecta anunțurile care reutilizează sau manipulează conținut deja identificat, inclusiv tactici adversariale, cum ar fi decuparea, rotația, ocluzia și noise aplicate imaginilor. Sistemele aplicate reclamelor sunt complet automatizate și crearea unei campanii publicitare care utilizează astfel de imagini manipulate va duce la respingerea automată de către sistem. ■

■ **Exemplu 5.3 — Utilizare SimSearchNet: Facebook MarketPlace.** Aceleași principii se aplică Facebook Marketplace, unde utilizatorii vând și cumpără mărfuri, unde Facebook a implementat "sute de modele de clasificare și detecție a obiectelor care funcționează bine în aceste condiții dificile din lumea reală" folosind Pytorch și recunoașterea imaginii Facebook cu Deep Learning pe hashtag-uri dezvoltată în 2018: https://engineering.fb.com/2018/05/02/ml-applications/advancing-state-of-the-art-image-recognition-with-deep-learning-on-hashtags. Ultimele abordări menționate pot clasifica imaginile nu doar prin analizarea imaginilor în sine, ci și prin utilizarea hashtagurilor adesea folosite de utilizatori ca și captions (subtitrări) la imaginile și anunțurile plasate pe piață. ■

Conceptele de SimSearchNet++ în folosite împotriva dezinformării prin identificarea imaginilor similare este, de asemenea, aplicat la conținutul de tip text: „când fact-checker-ii au identificat o informație falsă, dorim să identificăm copii ale acesteia chiar și atunci când au fost decupate sau modificate."

Sistemul utilizat pentru a detecta variațiile ale informațiilor identificate ca deja false folosește tehnologii precum ObjectDNA sau LASER, care este o platformă NLP open-source și care implementează zero-shot transfer al modelelor NLP de la o limbă la alta, suportând 90 de limbi, scrise în 28 alfabete diferite, utilizat pentru a evalua similaritatea semantică a propozițiilor.

R Politicile Pinterest legate de dezinformare medicală "Health misinformation"[2] interzice promovarea de tratatemente sau cure false pentru boli cronice sau termninale, precum și sfaturi legate de antivaccinare.

[2] https://policy.pinterest.com/en/community-guidelines

R Twitter nu permite media controlate de state (i.e. “state-controlled media”) pentru a-i utiliza produsele de publicitate. Provocarea este cum se realizează discriminarea dintre furnizorii publici de media independenți și cei controlați de stat.

IA este ubicuu în platformele media. De exemplu, în cazul Google:

- Google Photos: Funcționalitatea "Memories" utilizează învațarea automată pentru a sorta fotografiile relevante pentru o anumită temă
- Fotografii cinematice, învățarea automată este utilizată pentru a crea versiuni 3D ale fotografiilor utilizatorilor
- Gmail utilizează IA pentru: detecție de spam, clasificarea emailurilor in directoare diferite (i.e. bifurcation of emails), prezicerea textului in timpul redactării, detectarea răspusurilor automate (i.e. auto-reply detection)[a]
- Search: utilizarea procesării de limbaj natural (e.g. BERT) pentru corectarea gramaticii, înțelegerea paragrafelor, identificarea momentelor cheie în video.
- Soluțiile IA utilizate de Google Cloud sunt descrise în blog posts, dar nu apar prezentate clar in secțiunea de Termene și Condiții. Principiile Google legate de IA sunt enumerate la Google AI principles.

[a] https://technoitworld.com/how-does-gmail-use-artificial-intelligence/

Politicile Google

Google operează două servicii online importante care au un rol semnificativ în distribuirea și curarea știrilor la scară globală: Google Search și Youtube.

Google a recunoscut importanța marilor companii de tehnologie (Big Tech) în lupta împotriva știrilor false încă de la rezultatele alegerilor prezidențiale din SUA din 2016, atunci când Sundar Pichai, CEO al Google, a recunoscut într-un interviu la BBC că știrile false au afectat votul unor oameni[3]. De atunci, Google a investit în dezvoltarea nu doar a instrumentelor de combatere a știrilor false de pe platformele lor, ci și a finanțat inițiative precum Google Digital News Initiative (acum Digital News Initiative fund), care a introdus instrumente pentru a ajuta jurnaliștii să lupte împotriva știrilor false, și a oferit recent 25 milioane de euro (29.3 milioane de dolari) noului Fond European de Presă și Informare (European Media and Information Fund) pentru a combate știrile false.

Google Search rămâne principalul portal de acces la informații pe Internet pentru majoritatea utilizatorilor, având o cotă de piață de 86%. Google Search colectează și indexează conținut de la sute de miliarde de pagini web (https://www.google.com/search/howsearchworks/crawling-indexing/), doar indexul Google Search având doar o mărime de peste 100.000.000 gigaocteți. Procesul de indexare utilizează un instrument de descoperire a conținutului numit Knowledge Graph, care conectează obiecte și fapte, dar și relațiile dintre acestea, astfel încât orice interogare de căutare să poată furniza rezultate contextuale de înaltă calitate ale informațiilor relevante.

Deoarece conținutul online ce conține dezinformare este publicat pe web, șansele de a fi descoperit, colectat, indexat, și prezentat utilizatorilor este foarte mare. Astfel, consumul de informații false prin intermediul Google Search este o problemă cheie cu un impact considerabil.

Despre acest rezultat (About this Result). La 1 februarie 2021 Google a anunțat[a] funcționalitatea “About this Result” ca parte a soluției sale de căutare:. Această funcționalitate are scopul de a oferi mai mult context la căutarea online: “Cu acest context suplimentar, puteți lua o decizie mai informată cu privire la site-urile pe care doriți să le vizitați și ce rezultate vor fi cele mai utile pentru dvs.” Funționalitatea va oferi o descriere a site-ului sursă de pe Wikipedia, sau mai mult context suplimentar, cum ar fi data când

[3] https://www.bbc.com/news/business-37988095

Google a indexat prima dată site-ul respectiv.

[a] https://blog.google/products/search/about-search-results/

Efortul continuă o modificare din 2017 a algoritmilor de căutare Google pentru ierarhizarea site-urilor web și a elaborat noi directive pentru echipa de evaluare umană de peste 10,000 de persoane (https://blog.google/products/search/our-latest-quality-improvements-search/). Schimbarea a avut loc după ce un articol despre negarea Holocaustului a apărut în topul rezultatelor căutării în decembrie 2016. Algoritmii de clasificare a căutării au fost modificați utilizând feedback-ul de la evaluatorii umani, deși nu s-au oferit detalii cu privire la modul în care a fost utilizat feedback-ul pentru antrenarea algoritmilor de inteligența artificială sau modificări ale obiectelor conținute în Google Knowledge Graph. Modificarea a avut ca scop descurajarea conținutului de calitate scăzută în favoarea unor surse mai autoritative.

R Google oferă, de asemenea, instrumente pentru utilizatori și jurnaliști pentru a verifica informații, descoperi originea și gradul de încredere al imaginilor, detectarea de audio fals, profile de afaceri false, și chiar contribuții falsificate pe serviciul lor Google Maps.

Fact Check Explorer. Instrumentul Fact Check Explorer se adresează fact-check-erilor, jurnaliștilor și cercetătorilor, și constă într-o unealtă în care conținutul de la organizațiile de verificare a datelor este colectat și indexat, și un instrument de marcaj (Markup Tool) care permite să contribuția la baza de date de fact-checking datacommons.org disponibilă public prin intermediul unui flux de date.

Google Maps Google Maps a început să accepte coerntribuțiile utilizatorilor din 2010, cu peste 970 milioane de oameni contribuind cu comentarii, fotografii, evaluări, și informații factuale. Serviciul este, de asemenea, supus unor recenzii false care au ca scop scăderea sau creșterea reputației unei afaceri. Google a dezvoltat un algoritm ML în 2021, care scanează milioane de contribuții zilnice[a] și detectează și elimină conținutul care încalcă politicile sale, într-o varietate de limbi, combinat cu mii de operatori și analiști instruiți.

[a] https://blog.google/products/maps/google-maps-101-how-we-tackle-fake-and-fraudulent-contributed-content

Google My Business Vandalismul de conținut, click farms și operațiuni de recenzii false, sunt, de asemenea, vizate de algoritmul de învățare automată, în combinație cu politicile puse în aplicare pentru Google My Business[a]. Rezultatul este eliminarea sau revizuirea automată sau asistată uman de conturi și profilurile de afaceri, sistemele interne fiind responsabile de 85% din decizii.

[a] https://blog.google/products/maps/how-we-fight-fake-business-profiles-google-maps

Google News Lab, parte a Google News Initiative, este o echipa care susține jurnaliștii din întreaga lume cu training-uri și tehnologii[4] și are un mandat principal în combaterea dezinformării. Unele dintre lucrările realizate de Google News Labs sunt Trust Project, FirstDraft și Crosscheck, toate inițiative jurnalistice care vizează consolidarea credibilității mass-media și munca colaborativă împotriva dezinformării. Similar cu Facebook, Google News Lab cu partenerii săi se angajează în monitorizarea alegerilor, cu accent pe folosirea tehnologiilor existente puse la dispoziție de Google.

■ **Exemplu 5.4 — Suport pentru jurnaliști.** Un exemplu este inițiativa de a ajuta redacțiile din Asia să detecteze imagini false printr-un proiect comun al Google și Storyful susținut financiar de către Google Digital News fund, rezultând într-o aplicație numită The Source care utilizează tehnologia IA de la Google pentru a oferi acces instantaneu la istoria publică a unei imagini,

[4] https://newsinitiative.withgoogle.com/google-news-lab

permițând sortarea, analizarea și înțelegerea provenienței imaginii, inclusiv orice manipulare a acesteia[5]. ■

Abordarea You Tube

Principala platformă de distribuție și partajare a materialelor video de pe Internet este, de asemenea, vizată de dezinformare, un studiu publicat în British Medical Journal în mai 2020 arătând că mai mult de un sfert din clipurile video despre coronavirus de pe YouTube conțin informații înșelătoare sau inexacte[6].

Cu toate acestea, chiar și în contextul alegerilor din SUA din 2020, în cazul în care Facebook a ales să folosească rețeaua vastă 35.000 de fact-checkeri pentru a verifica conținutul și încetini distribuția conținutului de tip dezinformare în News Feed, YouTube a luat o abordare diferită, permițând utilizatorilor să folosească platforma chiar „dacă fac acest lucru în moduri care răspândesc conspirații nedovedite sau promovează afirmații false sau înșelătoare”[7]. YouTube se angajează să combată dezinformarea cu o serie de instrumente care elimină conținutul care încalcă politicile sale, favorizând surse de știri de încredere și reducând recomandările la conținutul de conține dezinformare la limită[8].

■ **Exemplu 5.5 — Aplicarea politicilor YouTube privind dezinformarea.**

- Informații medicale eronate (e.g. Covid19) cu privire la conținut care contrazice declarațiile Organizației Mondiale a sănătății (OMS)
- Practici înșelătoare, spam și înșelătorii sau impersonarea
- Discursul urii (hate speech).

R Determinarea a ceea ce este dezinformare se face printr-o combinație de sisteme și curatori umani[9]. Rezultatele evaluării umane sunt apoi utilizate pentru a îmbunătăți „sistemele de învățare automată bine testate pentru a construi modele care generează recomandări”.

La fel ca și în cazul căutării Google, YouTube are un motor puternic de recomandare personalizată, care poate livra conținut video pentru fiecare utilizator. Recomandarea video se face pe pagina de start și în secțiunea Up Next, folosind o combinație de personalizare (istoricul vizionărilor), performanță (atracție, angajament, satisfacție), și factori externi (interes subiect, competiție, sezonalitate)[10]. YouTube este însă criticat pentru favorizarea unor știri false și a unor clipuri video nepotrivite, potrivit unui studiu recent al Fundatiei Mozzila care arata că sistemul recomandă 71% din conținutul raportat de utilizatori[11]. Studiul solicită „legi de transparență de bun simț, o mai bună supraveghere și presiune a consumatorilor” – sugerând o combinație de legi care împuternicesc transparența în sistemele IA; protejează cercetătorii independenți astfel încât aceștia să poată interoga impactul algoritmic; responsabilizarea utilizatorilor platformei prin control robust (cum ar fi capacitatea de a renunța la recomandările „personalizate”) sunt ceea ce este necesar pentru a

[5] https://blog.google/around-the-globe/google-asia/new-tool-helping-asian-newsrooms-detect-fake-images/

[6] https://www.bbc.co.uk/newsround/52661709

[7] https://www.theverge.com/2020/11/12/21562910/youtube-2020-election-trump-misinformation-fake-news-recommendations

[8] https://www.youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation

[9] https://www.youtube.com/intl/en_us/howyoutubeworks/our-commitments/fighting-misinformation/#determining-misinfo

[10] https://blog.hootsuite.com/how-the-youtube-algorithm-works

[11] https://www.techtimes.com/articles/262534/20210707/youtube-recommendation-algorithm-false-information-inappropriate-videos.htm

reveni din cele mai grave excese ale inteligenței artificiale a YouTube[12]

■

Un raport publicat in 2019 [40] detaliază politicile folosite pentru a contracara știri false pe diferite servicii Google:

- Google Search, cu scopul de a face Internetul disponibil pentru toată lumea, conținutul nu este de obicei eliminat, dar numai deranked (declasat, deprioritizat), cu excepții foarte limitate.
- Google News este limitat la conținut de știri de calitate, cu un punct de vedere jurnalistic clar.

Furnizarea de context pentru utilizatori, utilizarea expertizei fact-checkerilor independenți, contra actorilor rău-intenționați și folosind algoritmi este combinația utilizată pentru a îmbunătăți calitatea conținutului și a filtra știri false. Documentul conține, de asemenea, o descriere a modului în care evaluatorii umani creează date de de antrenare pentru algoritmii IA ai Google.

Eforturile în construirea algoritmilor pentru YouTube sunt vizibile și în baza de date de cercetare a Google [13], cu peste 36 de publicații legate direct de YouTube. Unele dintre lucrări indică utilizarea Machine Learning și IA pentru a clasifica și recomanda conținut video: clasificare la scară largă a evenimentelor video, analiză acustică, modelare acustică cu rețele neuronale pentru transcriere video, clasificarea canalelor, analiza sentimentelor comentariilor, detectarea obiectelor, rețele neuronale pentru recomandări.

O descriere a algoritmului de recomandare YouTube este explicată referitor la tendința de a propaga conținut inflamatoriu vine de la cercetătorul Andre Ye[14]: „Cu toate acestea, în strălucirea sa, mulți au găsit o vină, în special în faptul că [algoritmul] aparent favorizează și propagă conținut inflamator, lăsând în praf creatorii de conținut onești." Explicația dată de Ye este că algoritmul tinde să se răspândească conținut șocant și inflamator, deoarece i se spune să optimizeze timpul de vizionare, nu aprobarea utilizatorului. Aceasta este echivalentul de clickbait pentru conținut scris.

Măsurile luate de Youtube sunt folosind fact checkeri pentru a filtra conținutul și mijloace pentru a diversifica atenția prin sugerarea de noi canale și recomanda videoclipuri.

Problema de a administra conținut video la scară largă a fost identificat de către Youtube în 2017, când videoclipuri destinate copiilor au inundat platforma cu parodii violente la desene animate populare pentru copii, cum ar fi Peppa the Pig sau Eroi în Pijama[15]. YouTube au lansat apoi o aplicație special concepută special pentru copii numiți Youtube Kids: „YouTube Kids este o versiune colorată, simplificată YouTube, plină de animații, culori strălucitoare și avatare de desene animate menite să-i mențină pe cei mai tineri utilizatori de internet implicați. Când utilizează aplicație, copiii pot vedea totul, de la compilații de cântece Nickelodeon la serii umor și clipuri video de gătit- un microcosmos vesel al YouTube. "[16]. Aplicația oferă conținut sigur selectat pentru copii, dar se concentrează în continuare pe timpul de vizionare și angajare și a fost criticată pentru că nu a putut dezactiva caracteristica autoplay a video-urilor.

R Un studiu publicat în 2019 care vizează detectarea automată a conținutului dăunător pentru copii a realizat „un clasificator capabil să distingă conținutul inadecvat care vizează copiii mici pe YouTube cu o precizie de 84.3% și să îl utilizeze pentru a efectua în premieră și pe scară largă caracterizarea cantitativă care dezvăluie unele dintre riscurile consumului de media YouTube de către copii mici" [79]. Studiul a susținut că YouTube este încă afectat de astfel de videoclipuri perturbatoare și că contra-măsurile implementate în prezent sunt ineficiente în ceea ce privește detectarea lor în timp util.

[12] https://techcrunch.com/2021/07/07/youtubes-recommender-ai-still-a-horrorshow-finds-major-crowdsourced-study

[13] https://research.google/pubs

[14] https://faun.pub/the-algorithm-worth-billions-how-youtubes-addictive-video-recommender-works-d75646dac6a3\T1\textquoteright0

[15] https://www.bbc.com/news/blogs-trending-39381889

[16] https://www.vox.com/recode/22412232/youtube-kids-autoplay

Perspectiva TikTok

Cunoscut în China sub numele de Douyin, este un serviciu de rețele sociale axat pe partajarea video, deținut de compania chineză ByteDance. Platforma social media este utilizată pentru a realiza o varietate de videoclipuri de scurtă formă, de la genuri precum dans, comedie și educație, care au o durată de cincisprezece secunde până la trei minute[17]. Platforma are aproximativ un miliard de utilizatori active lunar și este disponibil în 154 de țări și jumătate din utilizatori au sub 35 de ani.

Similar cu Youtube, TikTok folosește algoritmi de IA pentru a recomanda conținut video personalizat utilizatorilor, după cum este descris de specialistul AI Catherine Wong utilizând informațiile publice despre acest motor de recomandare al TikTok[18]. Acest motor de recomandare a conținutului a atras critici din partea publicului, acuzele principale fiind legate de dependența pe care o crează TikTok și de recomandarea de conținut nepotrivit copiilor.

Dependența utilizatorilor de TikTok TikTok pune la dispoziția creatorilor de conținut unelte de IA pentru realizarea de video-uri care să devină virale. Similar utilizatorilor Facebook și Instagram, utilizatorul TikTok petrece în medie 52 de minute pe zi în aplicație. În acest interval de timp, aceștia pot viziona peste 200 de videoclipuri, inclusiv anunțuri sau oferte direcționate cu atenție. O diferență între motorul TikTok și cele ale NetFlix, Youtube sau Facebook, este că TikTok nu prezintă utilizatorului o listă de recomdări de unde se poate allege următorul video de vizionat, ci decide automat ce va viziona utilizatorul. Această abordare bazată aproape integral pe IA oferă puțin control utilizatorilor, necompensată prin transparență legată de ce algoritmi sunt folosiți sau ce date ale utilizatorilor sunt învățate pentru generarea recomandărilor.

Recomandarea de conținut nepotrivit În China, Bytedance, compania-mamă a TikTok a plătit amenzi pentru conținut pornografic și reclame frauduloase. Compania a fost, de asemenea, acuzată că a colectat datele personale ale utilizatorilor cu vârsta sub 13 ani fără a lua acordul părinților, ceea ce a condus la o amendă de 5,7 milioane USD în SUA. Conform declarației Comisiei Federale pentru Comerț, compania a ales de bună voie să „urmărească creșterea afacerii chiar și în detrimentul pericolului copiilor” [a].

[a] https://www.nbcnews.com/tech/tech-news/tiktok-pay-5-7-million-over-alleged-violation-child-privacy-n977186

■ **Exemplu 5.6** Un articol Wall Street Journal detaliază experiența avută de un copil cu aplicația TikTok (https://www.wsj.com/articles/tiktok-algorithm-sex-drugs-minors-11631052944):

După ce s-a înscris pentru un cont pe TikTok, un utilizator de 13 ani a început să caute în aplicație „onlyfans” - numele unui site cunoscut pentru găzduirea de pornografie - apoi a urmărit câteva videoclipuri în rezultate, inclusiv două video-uri care vindeau pornografie. Apoi, utilizatorul a apelat la feedul personalizat „Pentru tine” (For You) al aplicației unde TikTok a livrat un șir de videoclipuri populare pe care mulți utilizatori le văd. Dar aplicația nu a uitat interesul tânărului utilizator pentru sex, servind rapid mai mult, inclusiv zeci de videoclipuri cu tentă sexuală. Pe măsură ce utilizatorul a parcurs videoclipurile care apar în feed, persistând pe cele mai orientate sexual în timp ce trecea mai repede pe lângă altele, feedul For You a fost în curând aproape în întregime dominat de video-uri TikTok care implică dinamica puterii sexuale și violența. Algoritmul aplicației a împins utilizatorul într-o gaură de iepure pe care mulți utilizatori o numesc „Kinktok”, cu bici, lanțuri și dispozitive de tortură. O parte din conținut este interzisă de platformă.

Contul a fost unul dintre zecile de conturi automate, sau roboți, creați de The Wall Street Journal pentru a înțelege ce arată TikTok tinerilor utilizatori. Acești roboți, înregistrați ca utilizatori cu vârste cuprinse între 13 și 15 ani, au fost creați pentru a răsfoi feedul TikTok For You, feedul foarte

[17] https://en.wikipedia.org/wiki/TikTok

[18] https://towardsdatascience.com/why-tiktok-made-its-user-so-obsessive-the-ai-algorithm-that-got-you-hooked-7895bb1ab423

personalizat, fără sfârșit, curatat de algoritmul de IA al platformei.

■

O analiză a videoclipurilor transmise acestor conturi a constatat că prin intermediul algoritmilor săi puternici, TikTok poate conduce rapid minorii - printre cei mai mari utilizatori ai aplicației - în nenumărate bule de conținut despre sex și droguri. TikTok a oferit un cont înregistrat ca un copil de 13 ani, cel puțin 569 de videoclipuri despre consumul de droguri, referințe la dependența de cocaină și metanfetamină și videoclipuri promoționale pentru vânzările online de produse de droguri și accesorii. Sute de videoclipuri similare au apărut în fluxurile celorlalte conturi minore ale Jurnalului.

De asemenea, TikTok le-a arătat utilizatorilor adolescenți ai WSJ mai mult de 100 de videoclipuri din conturi care recomandă site-uri pornografice cu plată și magazine de sex. Mii de alții provin de la creatori care și-au etichetat conținutul doar pentru adulți. Alții încă au încurajat tulburările alimentare și au glorificat alcoolul, inclusiv reprezentări despre băut și condus și jocuri de băut.

5.4 Reglementări împotriva deepfakes

Deepfake reprezintă manipularea conținutului audio sau video pentru a atribui unei persoane afirmații sau acțiuni pe care aceasta nu le facut, folosindu-se tehnici din inteligența artificială (e.g. rețele adversariale generative, autoencodere) [104].

Chiar dacă nu este menționată ca o problemă principală în propunarea de reglementare a IA, considerăm că referirile la tehnologiile IA de risc sunt aplicabile, fiind menționate tehnicile folosirii de bots și deep fakes (pag. 4). Capitoul 5.4 Transparency Obligations for Certain AI Systems se clarifică faptul că sistemele IA care impersonează sau manipulează conținut și interactionează cu oamenii, vor trebuie să fie marcate in acest sens (Utilizatorii unui sistem IA care generează sau manipulează conținut imagine, audio sau videocare seamănă semnificativ cu persoanele, obiectele, locurile sau alte entități sau evenimente existentești ar părea în mod fals unei persoane că este autentică sau veridică („fals fals”), trebuie sa semnalizeze daca materialul a fost generat sau manipulat artificial). De asemenea, este interzisă folosirea de sisteme IA destinate să denatureze comportamentul uman, prin care este posibil ca prejudiciile fizice sau psihologice (pag. 22), în această reglementare încadrându-se și conținutul online cu caracter de dezinformare sau “hate speech” creat și diseminat cu ajutorul IA.

R Platforme precum TikTok, Instagram, SnapChat includ opțiuni de manipulare a conținutului (e.g. filtre faciale, editare video), ceea ce: 1) creează un context social în care manipularea conținutului este încetățenită, 2) limitează sisteme de detecție a deepfake în a avea acces la conținut original, nemanipulat în vederea comparării acestuia cu conținutul generat artificial.

Riscurile asociate Deepfake au cel putin trei categorii [104]: psihologice (defăimare, intimidare, erodarea încrederii), financiare (extorcare, furt de identitate, fraue, manipulare acțiuni, descreditarea reputației) sau societale (manipulare media, manipulare alegeri etc). Introducerea tehnologie 5G și creșterea capacității de calcul vor face ajuta tehnologiile din spatele Deepfake pentru a produce conținut sintetic extrem de greu de distins față de conținutl real. Ca urmare riscurile percepute acum ca mici legate de deepfake, vor fi considerate mari în continuare.

Cadre de reglemetare la nivel UE relevante pentru deepfake (104)

- The AI regulatory framework
- The General Data Protection Regulation
- Copyright regimee
- e-Commerce Directive
- Digital services act

- Audio Visual Media Directive
- Code of Practice on Disinformation
- Action plan on disinformation
- Democracy action plan

Conform [104], opțiunile de reglementare depind de următoarele dimensiuni:

1. Reglementarea tehnologiilor. Abordarea bazată pe risc a UE [28] lasă loc de interpretări in acest domeniu - aplicațiile bazate pe deepfake nu intră în categoria de risc ridicat (sunt necesare doar cerințe minime legate de etichetare și semnalizare de interacțiune cu IA), dar ar putea intră în grupa aplicațiilor interzise [104].
2. Reglementarea generării de deepfakes. Această dimensiune se referă la reglementarea utilizatorilor de aplicații care pot genera deepfake.
3. Reglementarea distribuirii. Sunt vizate reguli pentru platformele media care sunt responsabile pentru distribuirea sau ștergerea unor astfel de creații sintetice. Astfel de reguli ar putea responsabiliza furnizorii de servicii digitale de a utiliza sistem de detecție, limitarea răspundirii sau eliminarea producțiilor deepfake.
4. Protecția victimelor. Dimensiunea apartine cadrului mai larg de reglementare a drepturilor personale.
5. Reglementarea audienței. Este o dimensiune importantă pentru limitarea riscului de redifuzare din partea celor care au intrat în contact cu astfel de producții sintetice.

Articolul 52 din propunerea UE (28) se referă la deepfakes: "users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated"

R Cerere si oferte pentru deepfakes (sau FakeNews 2.0) se tranzacționează în piețe dedicate.

R Există companii care furnizează servicii legate de deepfake: FaceApp a fost descărcată de peste 500 milione de ori, Wombo care aplică sincronizarea buzelor cu textul pentru a crea conținut satiric a fos tinstalată de mai mult de 10 milioane de utilizatori. Tehnologia deepfake fiind folosită la scară larga, o bună aparte din populația UE va fi obișnuită cu aceasta.

Strategii tehnice de prevenție

- Utilizarea și îmbunătățirea indicatorilor de autenticitate, e.g. eventual prin atribuirea de identificatori unici prin tehnologii precum blockchain sau distributed ledger [46] (deși nuărul de imagini, utilizatori ridică probleme de fezabilitate)
- Implementarea în cipurile camerelor pentru introducerea unui marker digital.
- Dezvoltarea de unelte pentru antrenarea persoanelor în detectarea conținutului sintetic.

R Sensity AI estimează că mai mult de 90% din producțiile deepfake conțin pornografie nonconsensuală [39] din care 90% dintre victime sunt femei. S-a avansat argumentul că [104] (pp. 34) ca astfel de unelte sunt "gendered–by-design" deorece sunt antrenate pentru a genera nuduri de femei.

■ **Exemplu 5.7 — Deepfakes pentru parlamentarii europeni.** Membrii ai parlamentului european a fost induși în eroare ca au o întâlnire video cu Volkov, șeful de campanie al lui Navalny, liderului opoziției în Rusia [90].

■

Detectarea DeepFakes la Facebook Videoclipuri fabricate sunt, de asemenea, vizate de echipa de Facebook AI RED) în colaborare cu Microsoft, Amazon Web Service și The PartnershiponAI.org. Aceste organizații lucrează împreună prin intermediul Deep Fake Detection Challenge (DFDC), un efort deschis și colaborativ pentru a crea noi tehnologii pentru detectarea deepfakes. DFDC a reușit să creeze un set de date[a] de peste 100,000 de videoclipuri cu opt algoritmi de modificare facială disponibile din iunie 2020.

n iunie 2021 cercetătorii Facebook în parteneriat cu Michigan State University (MSU) au anunțat un nou algoritm „de detectare și atribuire a DeepFakes care se bazează pe reverse engineering dintr-o singură imagine generată de IA la modelul generativ utilizat pentru a o produce"[b]. Abordarea utilizează un model de reverse engineering Machine Learning care depășește dependența modelelor anterioare, și care a necesitat un set de date suficient de mare pentru ar găsi un videoclip manipulat bazat pe originalul utilizat pentru instruirea modelului. Abordarea reverse engineering se bazează pe "descoperirea modelelor unice din spatele modelului IA folosit pentru a genera un singur deepfake", care poate identifica deepfakes generate folosind același model și care nu necesită să aibă un model învățat de mașină cu toate videoclipurile originale.

Nu este clar dacă algoritmii de detecție falsă profundă sunt utilizați de oricare dintre serviciile și produsele Facebook încă.

[a] https://ai.facebook.com/datasets/dfdc

[b] https://ai.facebook.com/blog/reverse-engineering-generative-model-from-a-single-deepfake-image

Detectarea DeepFakes la Google. Google face, de exemplu, și cercetare pentru detectarea de audio manipulat folosind tehnologia de sinteză audio utilizată de Google Maps, Google Translate, și Google Home (https://blog.google/outreach-initiatives/google-news-initiative/advancing-research-fake-audio-detection/). Această inițiativă, similară cu cea de detectare a DeepFakes a Facebook, are ca scop detectarea conținutului online manipulat care imită vocea unor oameni reali. Un set de date a fost pus la dispoziție în 2019, invitând „cercetătorii din întreaga lume să prezinte contramăsuri împotriva conținutului audio fals (sau „falsificat"), pentru a face sistemele de verificare automată a vocii (ASV) mai sigure". Detectarea vorbirii manipulate poate avea, de asemenea, un impact al sistemelor de identificare a vorbitorului utilizate de mulți furnizori, cum ar fi băncile, pentru autentificarea clienților pe baza amprentei vocale.

5.5 Unelte IA împotriva dezinformării online

Propagarea rapidă și necontrolată a dezinformării este facilitată în primul rând de mediul online, site-uri web, rețele sociale, platforme de distribuție conținut video, aplicații de mesagerie, si nu este de puține ori susținută de diferite grupuri de interes, chiar și de actori statali. Particularitățile fenomenului dezinformării în mediul online comparativ cu mediile tradiționale (presă scrisă, radio și TV) nu ține doar de potențialul de distribuție și audiența mult mai mare, ci și de actorii implicați in fenomen: creatori de conținut, distribuitori de conținut și consumatori. Dacă în cazul TV și radio responsabilitatea pentru conținut este a creatorilor (redacțiile), în mediul online distribuitorii de conținut au și rol de creatori de conținut, cu capacități de moderare, amplificare sau blocare a distribuției conținutului folosind o combinație de curatori umani și tehnologie, sau chiar facilitând amplificarea distribuției de conținut prin intermediul unor campanii plătite, platforma de distribuție având aici și rol de advertiser. Inteligența Artificială este în acest context și o oportunitate și o amenințare. Folosirea IA pentru colectarea, procesarea și analiza masivă de date nestructurate din mediul online pentru detecția automată a conținutului cu potențial nociv sau generat automat și suportul activității umane de analiză și verificare de informații (fact checking). Principalele tehnologii de IA folosite sunt cel din domeniul Procesării Naturale de Limbaj: sumarizare automată, traducere automată, comparație semantică, analiză de sentiment, clasificare, extragere de entități [4], speech to text, cât și algoritmi de identificare imagini si video manipulate deep fakes. Dar IA este folosită și pentru crearea de conținut sintetic, text, audio sau video (deepfake), cât și orchestrarea

amplificării de conținut pe rețelele sociale prin utilizatori-roboți (bots).

Chiar și în jurisdicțiile în care știrile fabricate nu au fost reglementate direct și legile pentru defăimare și calomnie sunt încă utilizate pentru a suprima în mod legal dezinformarea, principala provocare este abordarea cantităților imense de conținut online, care trebuie să fie analizate pentru a identifica conținutul de încredere la timp real, dar și pentru a monitoriza corectarea informațiilor. Astfel, problema dezinformării online reprezintă, de asemenea, o oportunitate de a stimula inovarea în crearea de abordări și tehnologii inovatoare.

O colecție de astfel de instrumente a fost colaționată de către RAND Corporation[19] „identificate [...] prin căutări online, articole care revizuiesc instrumente și progrese în acest domeniu, precum și discuții cu experți (de exemplu, cei implicați în elaborarea sau finanțarea de instrumente)". Cele 82 de aplicații din lista RAND a „instrumentelor care luptă împotriva dezinformării online" sunt clasificate în detectarea de bots/ spam, educație / formare, verificare, whitelisting, scor de credibilitate, trasabilitatea dezinformării, coduri, și standarde. Instrumentele de pe această listă care utilizează tehnici de inteligență computațională sunt doar câteva:

- FakerFact, soft care este în prezent scos din producție și a făcut uz de algoritmul de clasificare de învățare automată numit Walt pentru a identifica conținutul de tip opinie, satiră, și așa mai departe, analizând direct conținutul articolelor analizate.
- Mai multe instrumente abordează problema de bots, cum ar fi BotSentinel, Botometer și Hoaxy.

Lista RAND de instrumente[20] care conțin aplicații care nu mai sunt active sau întreținute, oferă o bună imagine de ansamblu asupra modului în care mediul academic, industria, sau freelancerii au abordat problema știrilor on-line false cu instrumente software, fie concentrându-se pe analiza de rețea și comportamentul inautentic pentru a detecta bots, pe construirea de baze de date cu surse de încredere, cu ajutorul fact-checkerilor umani, concentrându-se pe aspectul educațional, sau prin aplicarea IA la conținutul în sine pentru a extrage informații utile care ar ajuta utilizatorii să identifice dezinformarea.

Abordarea principală în utilizarea inteligenței computaționale pentru a identifica conținutul online de știri false este puternic axat pe învățare automată (Machine Learning - ML) și procesarea limbajului natural (Natural Language Processing - NLP), aplicată conținutului site-urilor web, spre deosebire de abordarea whitelisting / blacklisting, abordare care utilizează liste cu surse de încredere, care sunt apoi oferite utilizatorilor prin intermediul aplicațiilor web, aplicații mobile sau plugin-uri pentru browser-ul web. Acest tip de abordare (whitelists/blacklists) a fost criticată pentru că nu a reușit să țină pasul cu dinamica fenomenului, lipsa mecanismului de feedback și prejudecată politică potențială, un exemplu fiind extensia Chrome „B.S. Detector" care a primit o serie de recenzii nefavorabile în momentul când a fost lansată în 2016 pe platforma ProductHunt[21].

Lista de aplicații software care utilizează inteligența artificială pentru detectarea știrilor false și care sunt în producție este destul de scurtă:

Factmata Lansat în ianuarie 2017 în Regatul Unit de antreprenorul Dhruv Ghulati, după o finanțare a Google Digital News Initiative (DNI), Factmata a fost conceput pentru a verifica automat declarațiile pe social media, comentariile articolelor, conținutul articolelor, și orice text de pe Internet[a]. Acest software utilizează NLP pentru a detecta și clasifica automat conținutul online care ar putea fi dăunător sau nedemn de încredere. Facmata a dezvoltat o platformă dedicată de adnotare pentru conținut și au recrutat 24 comunități de specialiști și peste 2,000 de experți – inclusiv jurnaliști și cercetători – pentru a genera în mod activ un set de date de instruire, conform nesta.org.uk[b]. De la lansare, Factmata este acum poziționată ca o platformă care poate monitoriza conținutul de pe Internet și poate analiza riscurile și amenințările acestuia,

[19] https://www.rand.org/research/projects/truth-decay/fighting-disinformation/search.html

[20] Deși incomplete comparativ, de exemplu cu https://www.eu-startups.com/2020/03/10-european-startups-fighting-fake-news-and-disinformation

[21] https://www.producthunt.com/posts/b-s-detector

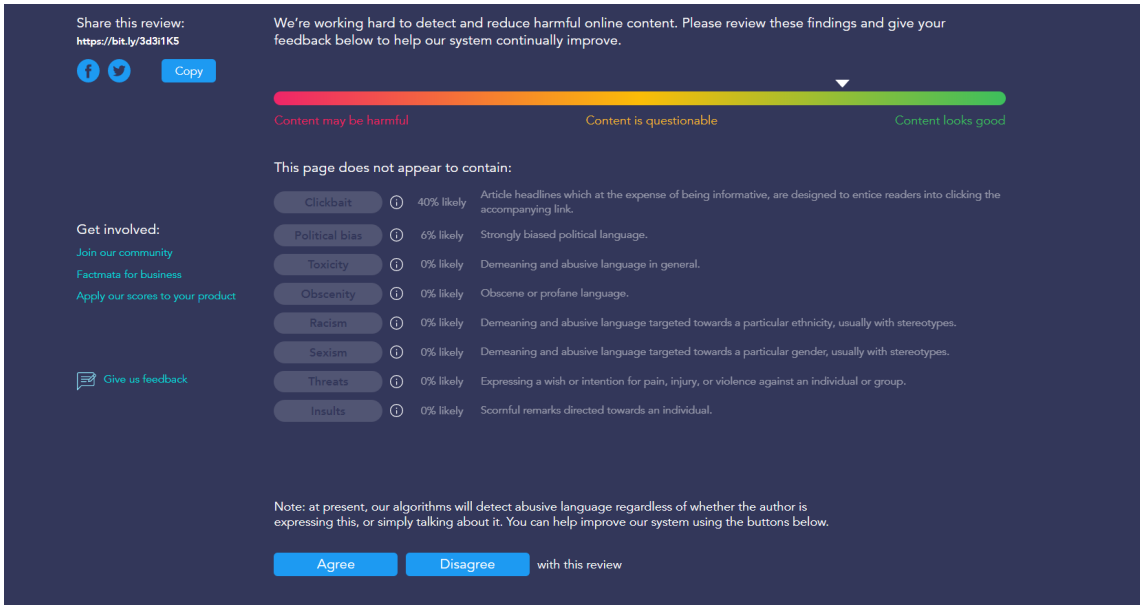

Figura 5.2: Exemplu de utilizare Factmata

axată pe monitorizarea narativelor și informațiilor despre consumatori și produse (Narrative Monitoring and Consumer and Product Insights), și deservesc clienți precum Taboola, Mediacorp și Silverbullet. Tehnologia IA utilizată de Factmata include analiza sentimentelor, clasificarea automată, similitudinea semantică, extragerea entităților denumite, detectarea poziției. Motoarele de clasificare automată sunt concepute pentru a acoperi o varietate de taxonomii, cum ar fi clickbait, conținut hiperpartizan, sexism, discurs al urii (hatespeech); amenințări, toxicitate lingvistică, obscenitate, insulte, subiectivitate, conținut emotiv, conținut controversat, și atitudine agresivă. Limbile suportate de aceast platformă de NLP este doar limba engleză. Factmata poate fi testată gratuit la `try.factmata.com` (Figura 5.2).

Factmata operează, de asemenea, extensia Google Chrome TrustedNews care măsoară obiectivitatea oricărui articol în mod automat, prin utilizarea unui model unic de învățare automată instruit de experți în domeniul științei datelorde la DataRhine, algorithm care dă fiecarei propoziție un scor: 0 în cazul în care propoziția este subiectivă, 1 în cazul în care este obiectivă (Figura 5.3). Acest algorithm este antrenat la nivel de propoziții (sentence level) folosind un set bogat de caracteristici, inclusiv expresii care poartă opinii (fraze pe care autorii le folosesc pentru a reflecta gândurile lor individuale, credințe, și atitudini) și alți indicatori semantici[c].

[a] `https://medium.com/factmata/whats-next-for-factmata-2df231bd6fe9`
[b] `https://www.nesta.org.uk/feature/ai-and-collective-intelligence-case-studies/factmata`
[c] `https://trusted-news.com`

TrustServista TrustServista este o platformă online de verificare a știrilor construită de start-up-ul românesc Zetta Cloud din Cluj-Napoca. Produsul a fost, de asemenea, finanțat de Google DNI, în etapa 1 a programului în 2016[a] și a folosit inițial motoare de analiză a textului de la compania americană Basis Tech, Rosette. De atunci, Zetta cloud și-a dezvoltat propria platformă multi-limbă NLP, utilizată de TrustServista și, de asemenea, oferită în comerț ca produsul distinct numit IntelliDockers (https://www.intellidockers.com/), care oferă o suită completă de capacități NLP pentru peste 50 de limbi, inclusiv est-European și Asian. Motoarele de procesare limbajului natural ale Zetta Cloud sunt motoare adaptabile, care pot fi create sau adaptate de utilizatori fără a avea cunoștiințe de Data Science sau Machine Learning, prin intermediul modului Factory. TrustServista este conceput pentru a verifica automat gradul de încredere al articolelor online prin analizarea conținutului site-ului web cu algoritmii NLP, construirea de clustere pentru știri, reprezentate ca grafice de relație semantică, și determinarea

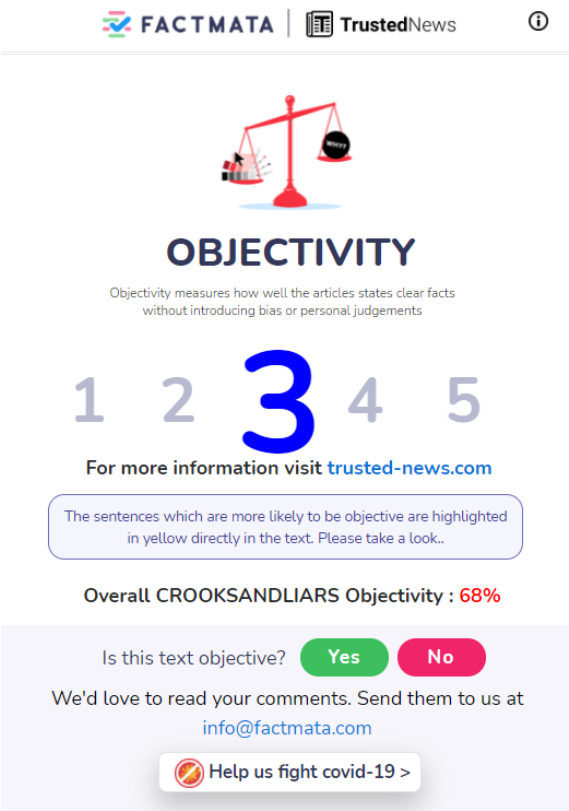

Figura 5.3: Exemplu de utilizare Factmata

originii informației, denumite Patient Zero. Conceptele utilizate de TrustServista sunt documentate în Raportul de Verificare Automată a Conținutului Digital publicat în 2018[b]. Platforma colectează surse de date online, cea mai mare parte feed-uri RSS și fluxuri ATOM, și apoi procesează fiecare articol cu motoarele sale NLP: Extragere de conținut, rezumare automată, analiza sentimentelor, extracție de entități, clasificare IPTC și IAB, și clickbait, generând un scor de încredere (Figura 5.4).

Toate articolele sunt apoi comparate între ele cu un motor de comparare similitudine semantică în mai multe limbi, pentru a determina grupurile de știri (clustere) și apoi pentru a identifica traseul de propagare a informației prin intermediul graficelor semantice (Figura 5.5).

TrustServista este disponibilă ca aplicație web, REST API-uri pentru integrarea cu alte sisteme IT, și o extensie Google Chrome[c]. Motoarele de procesare a limbajului natural (NLP) IntelliDockers care sunt integrate în TrustServista, pot fi adaptate de către clienți prin modului de no-code AI (IA fără a scrie cod software) Factory, lansată de Zetta cloud în 2021: [d]. O implementare recentă a TrustServista este prin intermediul platformei DO One Brave Thing (DOBT), un proiect finanțat de UE care vizează prevenirea radicalizării și a extremismului online. DOBT oferă un serviciu web care este construit folosind TrustServista pentru a determina dacă o știre este demnă de a fi citită sau distribuită online[e]. Serviciul funcționează pentru articole de știri în limba engleza, română, poloneza, maghiara, germana, și italiană.

[a] https://newsinitiative.withgoogle.com/dnifund/report/battling-misinformation/trustservista-question-trust

[b] https://www.trustservista.com/2018/04/white-paper-automated-approaches-to-digital-content-verification/

[c] https://www.trustservista.com/trustservista-extension/

[d] https://www.intellidockers.com/factory/index.html

[e] https://onebravething.eu/brave-web-service

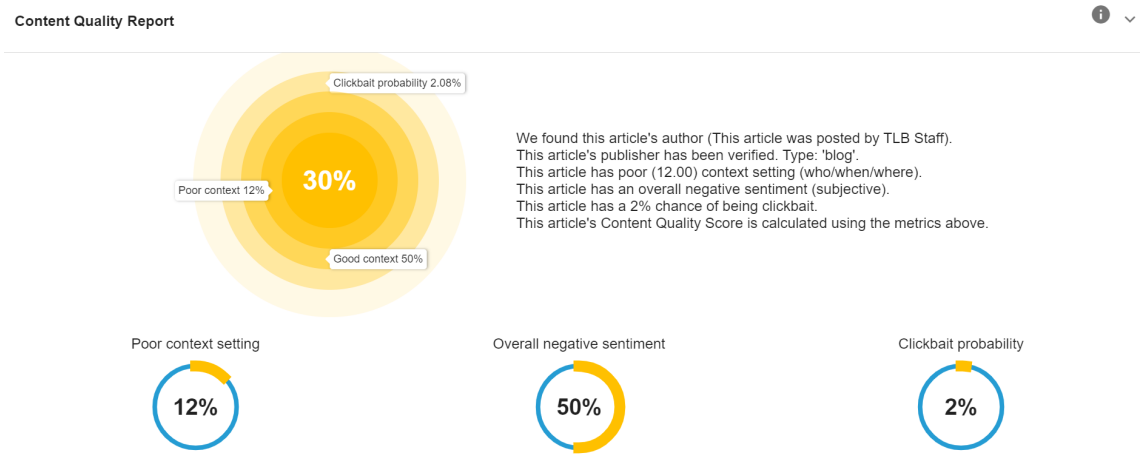

Figura 5.4: Exemplu de utilizare TrustServista

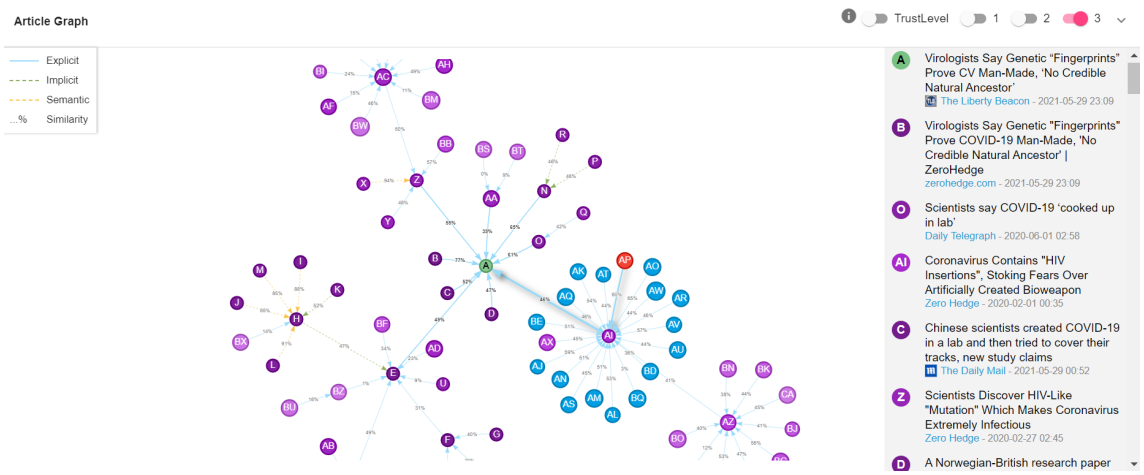

Figura 5.5: Exemplu de utilizare TrustServista

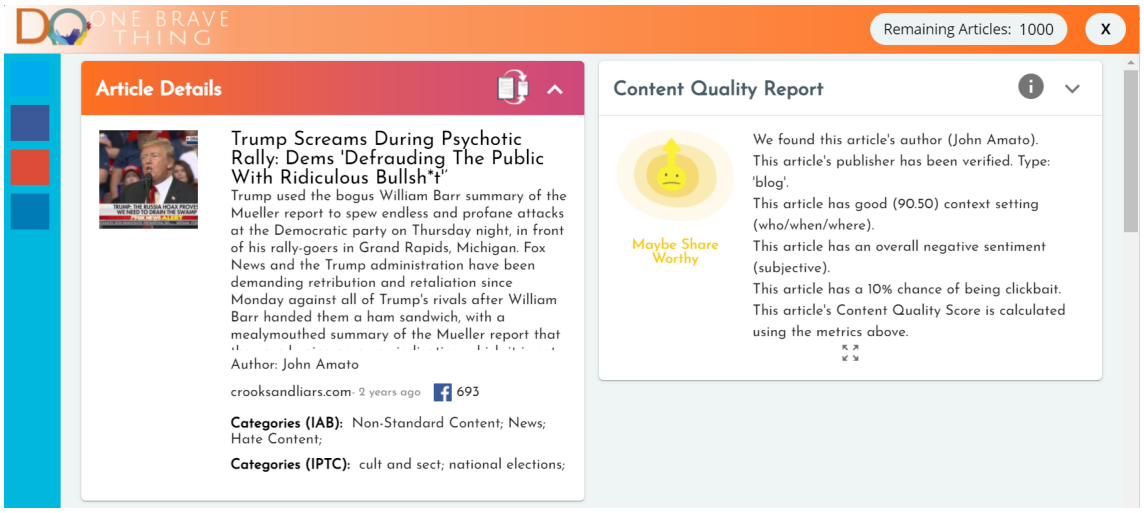

Figura 5.6: Exemplu de utilizare TrustServista

Fraunhofer Societatea Germană Fraunhofer, specializată în cercetarea științei aplicate, a lansat 2019 februarie un "software care poate detecta automat știri fabricate". Instrumentul utilizează învățarea automată pentru a analiza atât textul, cât și metadatele, axat pe Twitter și alte site-uri web și a fost construit folosind corpus de date provenit de la știri verificate cunoscute și știri false, deci cel mai probabil, folosind o abordare de clasificare automată. Conform comunicatului de presă, „pentru a filtra știrile false, cercetătorii folosesc tehnici de învățare automată care caută automat markeri specifici în texte și metadate", luând în considerare erorile lingvistice, punctuația incorectă, ortografia, verbul sau structura frazei. Instrumentul este, de asemenea, capabil să identifice hate speech. Nu se știe dacă instrumentul este utilizat în producție de către vreun client din sectorul public sau zona comercială.

Logically "the one in all threat intelligence platform", combină Inteligența artificială și experți în domeniul Open Source Intelligence (OSINT) pentru a investiga dezinformarea. Soluția este livrată ca aplicație web, iOS mobil și aplicație Android și o extensie de browser. Compania Logically a fost fondată în 2017 în Regatul Unit al Marii Britannii și a strâns fonduri în valoare de 2.77 milioane EUR în 2020.

Tehnologia utilizată de Logically este inteligența artificială (procesarea limbajului natural) și Knowldge Graphs pentru a extrage cuvinte cheie, subiecte și concepte, a clasifica conținutul, a identifica opinii, citate și evenimente și a determina polaritatea, prejudecata și toxicitatea conținutului[a].

Algoritmul de detectare a știrilor false este prezentat ca o evaluare automată a credibilității, construită pe o abordare cu trei axe a verificării - rețea, metadate și conținut - asemănătoare abordării comercializate de platforma TrustServista în 2018, inclusiv utilizarea conceptului „Patient Zero". În prezent, platforma acceptă numai limba engleză pentru procesarea conținutului (Figura 5.7).

Logically este, de asemenea, în curs de dezvoltare de instrumente de investigare multimedia pentru a identifica media manipulate, inclusiv falsuri profunde.

[a] https://www.logically.ai/tech?hsCtaTracking=b36e5c84-fb49-46ea-90a0-2b0d25d2559c%7Cc569dff6-3e15-46dd-8ef7-de870780bbbe

Buster.ai este o soluție dezvoltată în Franța în 2019 folosind algoritmi pentru a evalua credibilitatea conținutului, de la clipuri video la articole de știri. Buster.ai este oferit ca un portal web, REST API, și extensie de browser. Platforma poate audita informații înșelătoare și conținut multimedia prin autentificarea fețelor împotriva alterării Deepfake, detectarea conținutului fals generat de Inteligența Artificială și vizualizarea propagării informațiilor [11].

LiT.RL News Verification Lit.RL Browser (pronunțat "Literal") este un instrument de cercetare pentru consumatorii de știri, jurnaliști, editori, sau profesioniști intelligence, lansat în 2018. Instrumentul analizează limbajul utilizat în paginile web de știri digitale pentru a determina dacă acestea sunt clickbait, știri satirice, sau știri falsificate. Browser-ul lit.rl se bazează pe conceptul News Verification Suite al Victoria Rubin (Universitatea Western, Canada) [91].

Soluția este open source și livrată ca un web browser distinct construit pe baza Chromium. Detectorul de clickbait utilizează un algoritm care analizează legăturile găsite pe o pagină web și analizează titlurile știrilor. Motorul de detectare a satirei funcționează pe pasaje mai lungi de text și dă valoarea nominală ca procent. Cei trei algoritmi NLP lucrează pentru limba engleză și sunt algoritmi NLP de clasificare automată pentru clickbait (până la 94% precizie pe un set de teste de 5670 texte), satiră (până la 84% precizie pe un set de teste de 95 texte) și text falsificat (până la 71% precizie pe un set de teste de 28 texte).

Storyzy este o platformă franceză destinată siguranței brandurilor (brand safety), analizei informațiilor și datelor și alfabetizării în domeniul media digitală. Numele complet al platformei este Storyzy Database of Fake News sites. Storyzy oferă algoritmi pentru a detecta automat site-uri web, bloguri, și canale video bazate pe gradul lor de încredere și a clasifica aceste surse în categorii, suportând conținut în 7 limbi. Platforma descoperă și clasifică constant sute de surse noi de date, totalizând o bază de date cu peste 30.000 de surse de dezinformare (site-uri, bloguri, canale video).

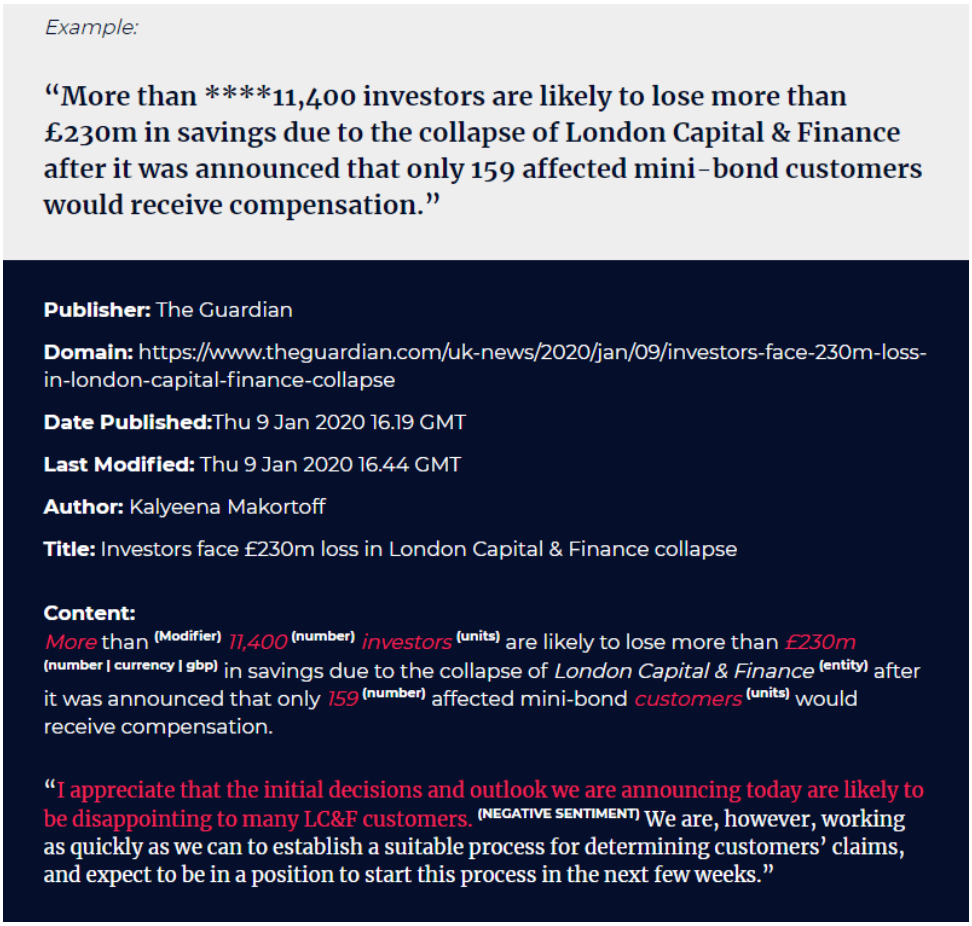

Figura 5.7: Exemplu de utilizare Logically

CLEF2021 CheckThat! Lab În cadrul conferinței CLEF2021, organizată la București [70] [69], laboratorul de verificare a știrilor fabricate a atras 132 participanți, evaluarea uneltelor realizându-se pe trei direcții:

1. Ordonarea unor tweet-uri pe baza check-worthy [96]
2. Ordonarea unor afirmații verificate in prealabil raportat la un tweet clasificat ca check-worthy [97]
3. Detectarea știrilor fabricate și clasificarea în patru clase: știri false, parțial false, adevarate, alte [98]

5.6 Propunere pentru combatarea dezinformării în România

Având în vedere că legislația în vigoare se adresează fenomenului dezinformării doar în mediul audiovizual, organ de reglementare fiind CNA, o soluție care se inspiră din masurile luate la nivel European, dar și cele din Germania, UK și Singapore, ar fi înființarea unui birou specializat pe dezinformarea din mediul online ca parte a CNA, și aflat în strânsă colaborare cu Autoritatea de Reglementare a Inteligentei Artificiale. Acest birou (pe model POFMA Singapore) ar identifica conținutul online relevant pentru România și cu impact/distribuție semnificativă, de genul site-urilor de știri sau a blogurilor, dar și a paginilor sau postărilor din rețelele sociale, ar implementa legislația existentă din audiovizual asupra acestor surse, ar aplica sancțiuni și ar monitoriza modul în care sancțiunile sunt aplicate (modelul Network Enforcment Act din Germania).

Totodată, acest birou ar fi responsabil și de crearea de bune practici în presa online, într-o strânsă colaborare cu mediul academic și cel de media, devenind un hub de training și validare ale organizațiilor de fact-checkers. Dar fiind faptul că domeniul online este mult mai greu de monitorizat decât cel audiovizual, o mare parte din uneltele folosite vor fi din zona de IA. Pentru a asigura corectitudinea și folosirea etică ale acestor unelte, cu principiul "human in the loop", biroul va colabora cu Autoritatea de Reglementare a Inteligenței Artificiale și va facilita mediului economic și celui academic implicarea în proiecte, fie prin atragerea de finanțări europene sau organizarea de licitații, pe modelul folosit de CNA de a contracta servicii de monitorizare trafic în audiovizual.

R Ar fi utilă crearea unor studii de caz de știri fabricate cu impact în România, pe linia abordării NATO pentru combaterea dezinformării.

Biroul va avea în componență și o unitate de interventie rapidă pe model britanic, unitate cu mai mulți stakeholderi și contributori (ministere, agenții) interesați, care va putea oferi un răspuns coordonat și strategic la fenomene de dezinformare cu potențial impact la nivelul siguranței naționale.

■ **Exemplu 5.8 — Dezinformare în programe de știri.** Infectari in showbizz, vedete care stau cu sufletul la gura ■

■ **Exemplu 5.9 — Dezinformare în spoturi publicitare.** Asociaţia „Preţuieşte Calitatea", pâine ambalată!" ■

Impact propunere

Propunerea nu vizează doar zona de reglementare și aplicare de sancțiuni pentru conținutul din mediul online, ci se dorește a fi un motor pentru dezvoltarea capacităților la nivel național de combatere a dezinformării, atât în mediul privat cât și în sectorul public și academic, mapate pe direcțiile strategice ale UE:

1. Îmbunătățeste transparența știrilor online, implicând o partajare adecvată și conformă cu confidențialitatea datelor despre sistemele care permit circulația lor online;Rolul regulator, implementat transparent, cu posibilitate de apel al celor implicați, conform legislației și fără potențial de abuzuri;

2. Promovează cunoștințele media și informaționale pentru a contracara dezinformarea și pentru a ajuta utilizatorii să navigheze în mediul digital; Autorități publice, mediul academic, ONG-uri vor pute conlucra în cadrul acestui birou pentru definirea continuă de bune practici și educație;
3. Dezvoltă instrumente care să permită utilizatorilor și jurnaliștilor să abordeze dezinformarea și să încurajeze un angajament pozitiv cu tehnologiile informaționale în evoluție rapidă; Companiile de IA din România vor putea dezvolta sisteme utilizate pentru combaterea fake news, utilizând experiența de lucru cu situații și date reale ale experților biroului;
4. Protejează diversitatea și sustenabilitatea ecosistemului mass - media european de știri, prin realizarea de statistici și studii obiective cu experți interni și externi asupra peisajului online și media din România;
5. Promovează cercetarea continuă asupra impactului dezinformării în Europa pentru a evalua măsurile luate de diferiți actori și a ajusta în mod constant răspunsurile necesare, colaborare activa între experții biroului, mediul academic și beneficiari din zona agențiilor de intelligence și securitate națională.

Coordonare. Interacțiuni cu parteneri

- ARIA
- Consiliul National al Audiovizualului
- Organizații (e.g. Clubul Român de Presă)

II Politici și instituții pentru IA în România

6. Autoritatea de Reglementare pentru IA

Viziune. O autoritate pentru monitorizare și reglementarea inteligenței artificiale etice și un model descentralizat în care centre de audit - private sau publice - au competențe de monitorizare, verificare și certificare în diferite subdomenii și tehnologii ale inteligenței artificiale.

Rezumat. Capitolul schițează funcțiile de bază, atribuțiile și drepturile generale ale unei autorități pentru reglementarea inteligenței artificiale. Conform Artificial Intelligence Act al CE, se dorește inființarea unei astfel de autorități in fiecare stat membru începând cu 2023. Capitolul identifică instituții cu care această autoritate va coopera pentru monitorizarea siguranței aplicațiilor IA și a asigurării dezvoltării IA etice.

Autoritatea de Reglementare pentru Inteligența Artificială (ARIA)[1] ar putea funcționa ca autoritate administrativă autonomă de reglementare și supraveghere în domeniul inteligenței artificiale[2]

6.1 Misiune și scop

Misiunea autorității se realizează prin protejarea drepturilor fundamentale ale cetățenilor și utilizatorilor de aplicații IA în contextul riscurilor introduse de sistemele bazate pe IA. ARIA va furniza expertiză tehnica pentru reglementarea sistemelor bazate pe IA, astfel încât reglementările introduse să contribuie la:

- accelerarea inovației în IA în România,
- asigurarea încrederii publicului în sistemele bazate pe IA (pe linia "AI made in Europe")
- aplicațiile IA dezvoltate în România să devină un produs de export
- protejarea valorilor UE și armonizarea cu reglementările UE

R Deoarece cele patru practici interzise la nivel UE [28] vizează inclusiv posibile acțiuni ale

[1] Denumiri alternative: (i) Autoritatea Națională pentru Inteligența Artificială; (ii) Autoritatea de Reglementare a Inteligenței Artificiale; (iii) Autoritatea Națională pentru Reglementarea Inteligenței Artificiale

[2] Această opțiune presupune o autoritate sub control parlamentar, dar pot exista și alte variante în funcție de decidentul politic.

guvernelor (e.g. identificare biometrică in timp real în spații publice, evaluarea comportamentului social), rămâne de stabilit caracterul politic al ARIA

În contextul noilor direcții vizate de UE, se propune:

1. Înființarea în 2023 a Autorității de Reglementare pentru Inteligența Artificială (ARIA)
2. ARIA va avea inițial și rolul de autoritate de notificare responsabilă cu instituirea procedurilor pentru evaluarea, desemnarea și notificarea organismelor de evaluare a conformității și pentru monitorizarea acestora.
3. Cadrul pentru reglementările IA va avea în centru conceptul de certificare IA [93].
4. Organismele de evaluare a conformității vor certifica sistemele bazate pe IA aflate în grupa de risc ridicat sau a altor sisteme bazate pe IA pentru care existența unui certificat asigură incredere sau avantaj competitiv
5. ARIA va fi organizată astfel încât să nu existe conflicte de interese cu organismele de evaluare a conformității și să asigure obiectivitatea și imparțialitatea proceselor de evaluare a conformității. ARIA nu va oferi servicii de evaluare a conformității.
6. ARIA va reglementa funcționarea *spațiilor de testare în materie de reglementare a IA*. Spațiile de testare în materie de reglementare reprezintă zone unde reglementările sunt limitate și favorabile pentru testarea aplicațiilor bazate pe IA). Spațiile de testare în materie de reglementare a IA reprezintă un element important în facilitarea introducerii noilor tehnologii și a produselor inovative pe piață.
7. ARIA va elabora ghiduri pentru dezvoltarea de aplicații IA mai transparente, predictibile și verificabile
8. ARIA va furniza fonduri/granturi pentru start-upuri/experți pentru implicarea în procesul de standardizare
9. ARIA va organiza programe de pregătire și acreditare a auditorilor specializați în certificarea sistemelor bazate pe IA
10. ARIA va putea furniza (o lista de) experți în IA pe diferite subdomenii
11. ARIA va monitoriza și reglementa noile drepturi ale cetățeanului în contextul interacțiunii cu sisteme bazate pe IA, e.g. Dreptul la explicație, Dreptul la informare în cazul interacțiunii cu IA
12. ARIA va promova și rula portalul Romanian Data Space (pe linia European Data Space)
13. ARIA va organiza consultări publice înainte de a adopta reglementari, oferindu-se astfel părților interesate posibilitatea de a formula opinii și de a transmite observații asupra măsurilor propuse

Crearea spațiului de date la nivel național Pentru sprijinirea dezvoltării IA la nivel național, un prim pilon care trebuie asigurat este crearea spațiului de date la nivel național. Datele diferite trebuie tratate diferit: Datele din sectorul public (e.g. statistici, date despre mediu, mobilitate) trebuie să fie disponibile și utilizate pentru economie și populație. Deoarece nu toate datele pot fi publicate deschis, trebuie definit cadrul legal pentru partajarea datelor între părți. Politicile și contractele de utilizare a datelor trebuie să specifice cine are drept de acces și în ce scop pentru ca beneficiile prin utilizarea datelor să fie maximizate pentru societate. Cadrul legal trebuie să asigure condiții furnizorilor și utilizatorilor de date pentru exploatarea datelor intersectorial. Este de asemenea necesară reglementarea intermediarilor care se ocupă cu stocarea și asigurarea accesului la date pentru utilizarea acestora în aplicații de inteligență artificială. Cadrul legal stabil este cel care poate asigura încrederea între furnizorii și utilizatorii de date.

ARIA va avea resurse computaționale pentru:

- Gestiunea incidentelor cauzate de sisteme bazate pe IA,
- Sprijinirea inovării prin accesul democratic al unor startupuri în domeniul IA

Catalogul incidentelor provocate de aplicații IA AIID este o bază de date a incidentelor în care sunt implicate sisteme IA, fiind un proiect al organizației Partnership for AI. AIID conține mai mult de 1.000 de incidente [65], fiecăruia fiindu-i asociate mai multe rapoarte din presă. Catalogul urmează practicile din industria aviatică legate de menținerea unei baze de date a incidentelor și accidentelor aviatice, practicile din securitate cibernetică prin expunerea publică a peste 140.000 de vulnerabilitați (i.e. Common Vulnerabilities and Exposure),

ARIA va avea expertiza si resursă umană pentru:

- Acreditarea Centrelor de Audit
- Gestionarea (i.e. înțelegere și intervenție) în colaborare cu alte instituții a unor incidente sau situații cauzate de sisteme bazate pe IA (e.g. companii de dezinformare prin știri fabricate). ARIA va elabora planuri de contingență (interventie) pentru astfel de situații
- Formarea și acreditarea auditorilor specializați
- Achiziții de sisteme bazate pe IA în administrația publică
- Dezvoltarea de aplicații utile in monitorizarea aplicațiilor bazate pe IA din administratia publica (e.g. Sistem pentru Evaluarea Impactului Algoritmilor – similar cu Algorithmic Impact Assessment Tool

Spațiu de testare în materie de reglementare a IA (e.g. regulation sandbox) Spațiile de testare în materie de reglementare a IA creează un mediu controlat pentru testarea tehnologiilor inovatoare pentru o perioadă limitată de timp, pe baza unui plan de testare convenit cu autoritățile competente

■ **Exemplu 6.1 — Spațiu de testare în materie de reglementare a IA.**

- Zonă pentru testarea sistemelor autonome pentru livrare produse
- Zonă pentru testarea vehiculelor autonome (e.g. drone)
- Spații de date

■

■ **Exemplu 6.2 — Dreptul la explicație.**

- Refuzul din partea administrației publice de a acorda ajutor social unei persoane pe baza calculelor sau recomandării unui sistem bazat pe IA
- Refuzul de a primi viză pe baza recomandării unui sistem bazat pe IA
- Refuzul din partea administrației locale pe baza recomandării unui sistem bazat pe IA de a aproba organizarea unui eveniment
- Refuzul din partea sistemului juridic de a aproba eliberarea condiționată pe baza unei evaluări bazate pe o aplicație IA

■

6.2 Funcţiile de bază, atribuţiile şi drepturile generale

Funcţiile de bază, atribuţiile principale şi drepturile generale preconizate sunt descrise în continuare.

Obiectivele ARIA

1. Dezvoltarea inteligenței artificiale în România
2. Armonizarea legislatiei interne cu reglementările UE cu privire la IA, conform standardelor "AI made in EU"
3. Garantarea respectării de către dezvoltatorii de soluții IA a obligatiilor ce le revin în ceea ce privește transparența, confidențialitatea, robustețea sistemelor dezvoltate
4. Informarea dezvoltatorilor și a utilizatorilor de soluții IA cu privire la potențialele riscuri sau drepturi care ar putea fi încălcate

Drepturile ARIA

1. Să interzică importul, comercializarea şi utilizarea aplicațiilor bazate pe IA care nu au certificatul de siguranță conform grupei de risc definite în Regulamentul UE, a standardelor naționale, reglementărilor tehnice sau a altor acte tehnico-normative
2. Să blocheze în regim de urgență comercializarea și utilizarea unor "jucării smart" care sunt susceptibile de a avea breșe de securitate a celor care pot afecta negativ dezvoltarea copiilor sau a celor care stochează date colectate de la copii (e.g. conversatii, poze), asistenți virtuali pentru copii [a]

[a] Într-o investigație a BBC asistenții proiectați pentru interacțiunea cu copiii pe probleme de sănătate mentală au eșuat în a semnala abuzuri sexuale, i.e. Child advice chatbots fail to spot sexual abuse

Activitățile ARIA includ:

1. Realizarea cadrului juridic şi elaborarea reglementărilor specifice domeniului inteligenței artificiale
2. Autorizarea centrelor publice sau private de audit pentru sisteme bazate pe inteligența artificială
3. Monitorizarea sistemelor bazate pe IA care funcționează în administrația publică.
4. Crearea de registre publice cu sistemele de decizie automate care funcționează în administrația publică.
5. Stabilirea procedurii de contestare in cazul deciziilor luate pe baza unei aplicatii cu IA
6. Gestiunea incidentelor și a situațiilor de criză cauzate de sisteme bazate pe inteligență artificială
7. Participarea ARIA în organizațiile de standardizare
8. Corelarea cu domenii conexe (e.g. privacy, protecția datelor, securitate) pentru eficientizare, claritate, evitarea dublei reglementări, sau pentru evitarea posibilelor conflicte și ambiguități
9. Sprijinirea dezvoltării IA în România

R Sistemele de decizie automate din administrația publică vor avea scorurile pentru Algorithmic Impact Assessment (AIA) și Human Rights Impact Assessments (HRIA) făcute publice. De asemenea, descrierea scopului, explicații ale modelului de funcționare și informații despre dezvoltator. Informațiile vor fi publice, inclusiv în format structurat (i.e. linked data). Achizițiile de sisteme bazate pe IA in sectorul public, vor fi condiționate, printe altele, de existența evaluărilor de tip AIA și HRIA, conform standardelor naționale.

6.3 Principii de reglementare

Următoarele principii de reglementare a inteligenței artificiale au fost adaptate din [37] în scopul creionării unor linii directoare pentru ARIA:

Încredere in aplicațiile IA. Acceptarea aplicațiilor IA este condiționată în mod semnificativ de încrederea și validarea la nivel public. Reglementările de tip limitare sau facilitare trebuie adaptate nivelului de risc asociat fiecărei aplicatii. ARIA trebuie să comunice toate asumpțiile și incertitudinile legate de rezultatele preconizate – atât pozitive cât și negative – legate de reglementările propuse sau de certificarea sistemelor IA.

Participare publică. ARIA trebuie să ofere actorilor implicați oportunități de participare pe parcursul procesului de reglementare. ARIA este responsabilă de crearea de documente informative și de popularizarea standardelor. ARIA asigură oportunități previzibile pentru ca actorii interesați: i) să funizeze feedback pe documentele în lucru; ii) să evalueze dovezile disponibile și documentul de analiză a necesității reglementării respective.

Integritate științifică și calitatea informației. Informațiile furnizate de ARIA trebuie: 1) să respecte standardele de integritate științifică pentru informarea factorilor de decizie, 2) sa asigure încrederea publică în IA. Pentru acest demers, bunele practici includ menționarea transparentă a punctelor tari, punctelor slabe, beneficiilor preconizate, atenuarea biasului. ARIA se asigură că aplicațiile IA sunt predictibile, iar datele utilizate în procesul de antrenare au calitate suficientă pentru utilizarea vizată.

Evaluarea și gestiunea riscurilor. Aplicațiile IA vor fi împărțite în patru grupe de risc, conform abordării UE [28]. ARIA evaluează care riscuri sunt acceptabile și care aplicații prezintă riscuri de daune sau prejudicii inacceptabile sau pentru care costurile preconizate depășesc utilitatea asteptată. Evaluarea riscurilor se face transparent și continuu, la intervale de timp adecvate, pentru a lua în considerare informatiile și dovezile nou apărute.

Beneficii si costuri. ARIA analizează costurile sociale, beneficiile și efectele laterale înainte de introducerea unei reglementari. La analiza costurilor și beneficiilor, autoritatea consideră dependentele sistemului bazat pe IA, atât cele tehnologice (e.g., disponibilitatea și calitatea datelor), cât si factorul uman asociat implementarii sistemului IA în diferite industrii. Sistemul IA va fi comparat cu sistemul înlocuit, daca acesta există, de exemplu se compară probabilitatea de eroare a sistemului IA cu cea a sistemului existent. Daca sistemul IA nu înlocuiește un alt sistem, se evaluează riscurile și costurile neimplementarii sistemului IA. Daca beneficiile unei reglementari nu justifică costurile preconizate, ARIA poate identifica si promova abordari de tip reglementări nonjuridice: i) ghiduri de bune practici pe domeniu (e.g. playbooks, lessons learned), ii) programe pilot și experimente (e.g., tech sprints), iii) recomandarea de standarde (elaborate de industrie).

Flexibilitate. Abordarile pentru reglementare trebuie sa fie flexibile pentru a putea reacționa la schimbările tehnologice în IA. ARIA se asigură continuu ca reglementările în vigoare nu introduc dezavantaj competitiv firmelor din România. ARIA se asigură continuu ca lipsa unor reglementări nu introduce dezavantaj competitiv firmelor din România. ARIA va evalua retrospectiv reglementările introduse și va iniția măsuri de dereglementare. La introducerea unei reglementări, ARIA trebuie să justifice de ce lipsa reglementării blochează inovarea.

Nediscriminare. IA poate reduce discriminarea cauzată de subiectivitatea agentului uman, dar poate introduce noi tipuri de discriminare. ANRIA va evalua discriminarea introdusa de sistemul IA comparativ cu discriminarea existent.

Comunicare și transparență. Înainte de propunerea unei noi reglementări legată de creșterea transparentei, ARIA va analiza daca reglementările și normele în vigoare nu sunt suficiente.

Siguranță și securitate. Reglementările trebuie să încurajeze aspecte legate de securitate și siguranță pe intreg procesul de proiectare, dezvoltare, instalare și rulare (e.g., actualizare software) a aplicatiilor IA. Reglementarile asigura confidentialitatea și integritatea informatiilor procesate, stocate și transmise de aplicatiile IA. Reglementările trebuie să prevină riscuri de securitate cibernetică și a tehnologiilor adversariale (e.g. GAN) de utilizare IA.

Coordonare și cooperare. ARIA va putea fi eficientă doar prin colaborare extensivă la nivel național și UE cu parteneri strategici, colaborare cu industria, politici de stimulare, agenții de standardizare.

6.4 Coordonare și cooperare

În fazele inițiale, eficiența și utilitatea socială și economică a ARIA va depinde de modul în care va utiliza și încorpora expertizele acumalate în timp de alte agenții. Mecanismele de cooperare cu alți actori sunt deci deosebit de importante.

ARIA ar trebui să coopereze cu toate organismele naționale a căror arie corespunde aplicațiilor IA cu grad ridicat de risc, conform Anexei III din propunerea de regulament a UE [28]. De exemplu (Figura 6.1), la nivel național ARIA ar putea coopera cu:

Agenții de standardizare pentru elaborarea unui itinerariu de standarde în IA) (e.g. Organismul Naţional de Standardizare – ASRO, de exemplu pentru siguranța jucăriilor care includ aplicații bazate pe IA, i.e. Jucării electrice. Securitate)

Consiliul Național al Audiovizualului pe domenii precum știri fabricate de sisteme bazate pe IA, filtrare conținut pentru copii

Figura 6.1: Exemple de cooperare a autorității de reglementare a IA cu alte agenții

Protecția consumatorului în cazul manipulării consumatorilor de către algoritmi de generare recenzii false, algoritmi de stabilire a pretului.

Agenția Națională a Medicamentului și a Dispozitivelor Medicale din România - pentru monitorizarea sau certificarea sistemelor bazate pe IA instalate pe dispozitive medicale

Ministerul Muncii și Protecției sociale, Ministerul Educației, Ministerul Economiei pentru stabilirea strategiei naționale pentru dezvoltarea competențelor (similar cu National Skills Strategy în Germania)

Autoritatea Națională pentru Drepturile Persoanelor cu Dizabilități, Copii și Adopții pe linia jucăriilor cu IA, asistenților voce IA, jocurilor video, IA emoțional

Centre de audit pentru acreditarea și monitorizarea acestora

La nivel UE și internațional, ARIA ar putea coopera cu:

Agenții de furnizare a datelor precum European Data Space (e.g. The European Science Cloud - EOSC)

Agenții naționale pentru reglementarea IA e.g. German Observatory for Artificial Intelligence,

UNICEF pentru reglementarea jucariilor cu inteligență artificială

Ce poate sa facă sistemul universitar:

1. Dezvolte bune practici și standarde pentru IA
2. Asigură competențe pentru dezvoltarea, înțelegerea și utilizarea IA
3. Conștientizează agenții economici șî administrația cu privire la noile riscuri și tehnologii
4. Semnalizează pe baza dovezilor științifice cazurile în care dezvoltatorii de IA nu respectă normele etice cu privire la IA
5. Promovează parteneriate cu administrația publică pentru dezvoltarea de competențe pentru utilizarea aplicațiilor bazate pe IA.
6. Sprijină introducerea de laboratoare de IA pentru învățământul liceal

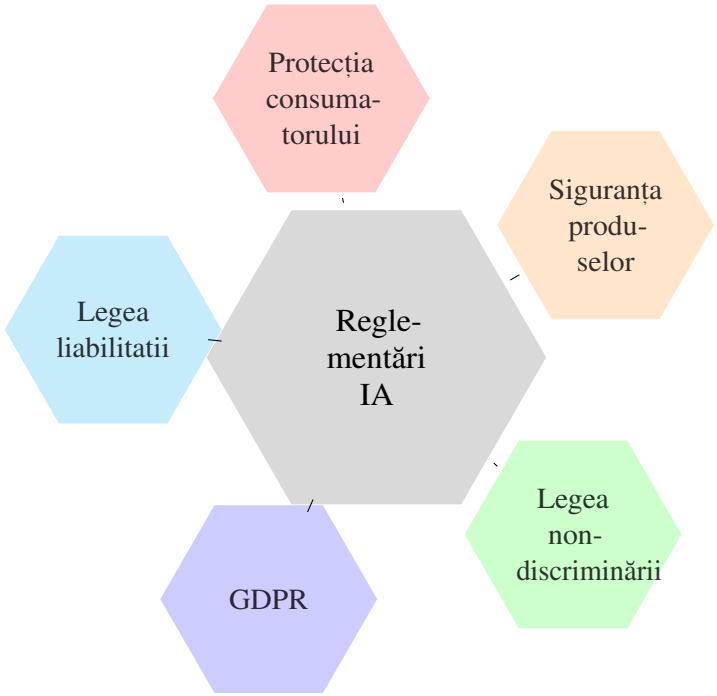

Figura 6.2: Reglementările pe IA vor trebui potrivite reglementărilor conexe existente

R Este o chestiune de timp până când elemente de IA vor pătrunde în curricula liceală: protejare împotriva dezinformării, etica datelor, interacțiunea cu asistenți virtuali, competiții.

R ARIA va avea obligația de a se afilia la organizațiile relevante în domeniu cu scopul de a: i) acumula expertiză și eficientiza activitatea, ii) promova interesele ecosistemului IA din România iii) facilita investițiile în IA in România prin cultivarea și promovarea etichetei "IA innovated in România".

Instituții similare: German Observatory for Artificial Intelligence Instituția asigură dialogul între guvern, mediul științific, mediul de afaceri și cetățeni. Urmărește atenționarea actorilor, fie dezvoltatori sau utilizatori de aplicații IA cu privire la constrângerile etice și legale ale IA. Are atribuții legate de identificarea reglementarilor oportune, crearea de spații de testare în materie de reglemntare a IA (e.g. living labs pe IA, stabilrea de centre pilot pentru IA (i.e. Centres for the Future), suport pentru persoanele fizice care dezvoltă aplicații IA (i.e. House of the Self-employed) sau monitorizarea ratei de pătrundere a IA în industrie.

GlobalPolicy.ai sprijină cooperarea interguvernamentală pentru IA.

IA pentru SDG (Sustainable Development Goals) Agenda pentru dezvoltare sustenabilă pentru 2030 adoptată de către UN în 2015 stabilește 17 obiective: 1) eliminarea sărăciei; ii) zero foame, iii) sănatate iv) educație de calitate, v) egalitatea genurilor, vi) apă curată, vii) energie curată și afordabilă viii) muncă decentă și creștere economică; ix) industrie, inovatie și infrastructură, x) reducerea inequalității xi) orașe și comunități sustenabile xii) consum și producție responsabile, xiii) acțiuni climatice, xiv) viată sub apă, xv) viața pe uscat, xvi) pace, justiție și instituții puternice, xvii) parteniate pentru aceste obiective. IA va juca un rol cheie în iv) educație de calitate, xi) orașe și comunități sustenabile.

6.5 Impactul reglementărilor

R Reglementările trebuie să urmărească asigurarea încrederii în aplicațiile IA. Fără încredere se restrâng domeniile de aplicabilitate dar și cantitatea datelor partajate. Fără date nu va fi combustibil pentru dezvoltarea aplicațiilor bazate pe învățare automată și implicit nu vor apărea firme și inovație pe această linie.

Impactul reglementărilor:
ARIA va promova reglementări care reduc barierele de dezvoltare și utilizare ale aplicațiilor IA: acces la spațiul european al datelor, open linked data

■ **Exemplu 6.3 — Impactul reglementărilor - Date deschise.** Reglementări care impun administrației furnizarea datelor deschise și norme de aplicare care specifică explicit modalitatea de actualizare a protocolelor de publicare sau granularitatea datelor (data granulare vs. agregate), pot avea impact imediat asupra dezvoltării aplicațiilor IA în România. ■

■ **Exemplu 6.4 — Impactul reglementărilor - Principiul "once-only".** Un exemplu de reglementare cu IMPACT asupra digitalizării administrației publice rămâne legea pe baza principiului "once-only" cu articol unic "administrației îi este interzis să ceară o informație de două ori." ■

Prin analogie cu principiul digitalizării once-only, se poate lua în considerate principiul prin care un set de date rezultat din procesarea cu un motor de IA (clasificare, extragere meta-date, etc) realizat de către o instituție publică, nu va trebui să fie procesat încă o data fără un motiv plauzibil. Implementarea acestui principiu poate duce la crearea unei baze de date instituționale comune, care să stocheze conținut și meta-date generate de IA disponbile pentru toate autoritățile și agențiile interesate. Un motiv plauzibil pentru reprocesarea datelor este disponibilitatea unui algoritm de IA mai avansat, care ar putea produce rezultate mai bune.

■ **Exemplu 6.5 — Acțiuni de stimulare a inovării - Civic Innovation Platform.** Urmărește sprijinirea cu fonduri a celor care doresc să dezvoltate aplicații IA. ■

Dezvoltarea ecosistemului IA. Un program eficient din punct de vedere al costurilor și beneficiilor lansat de ARIA ar putea viza sprijinirea persoanelor fizice (e.g. freelancer) care dezvoltă aplicații IA.

Dezvoltarea ecosistemului IA. ARIA va iniția scheme de susținere a celor care partejează datele. ARIA va susține financiar dezvoltarea de standarde în IA.

După identificarea problemei se decide dacă sunt necesare reglementari juridice sau sunt suficiente cele cu caracter nejuridic (e.g. ghiduri, standarde, norme). Se evaluează reglementările existente, dar și acțiunile potențiale ale agenților economici în contextul noilor reglementări vizate.

Justificarea unei reglementări
Justificarea introducerii unei reglementări trebuie să includă:
1. descrierea problemei,
2. scopul vizat (e.g., reglarea pieței, protejarea unor drepturi și libertăți, prevenirea discriminarii, protejarea securității naționale, sprijinirea companiilor IA din Romania

La introducerea unei reglementări se compara beneficiile sistemului IA cu sistemul vizat a fi inlocuit. Beneficiile și costurile trebuie cuantificate atât în valoare monetară, cât și în unități fizice (e.g. număr de accidente evitate, timp mediu câștigat). Analiza trebuie să se bazeze pe cele mai bune informații științifice, tehnice și economice.

R Tipul de reglementare – juridică, nonjuridică, fără reglementare, limitare, favorizare – este

decis de clasa de risc asociată aplicației IA.

R ARIA trebuie să asigure oportunități previzibile pentru actorii interesați: i) drafturi publice, ii) întâlniri publice, iii) șabloane pentru generarea documentației în vederea certificării.

R ARIA organizează evenimente pentru acceptare publică a sistemelor IA (similar cu, e.g. European Robotics Week: public meets the robots.

R ARIA reglementează și informează publicul cu privire la predicții făcute de IA care nu au fundament științific sau relevanță statistică (e.g., predicția succesului la un loc de muncă pe baza tonului vocii sau a microexpressilor, ordonarea CV-urilor pe baza unor metrici care includ și lungimea CV-ului).

R ARIA reglementează nivelul de expertiza necesar personalului care interacționează cu sistemele IA.

R ARIA va promova proiectarea hardware-software a aplicațiilor bazate pe IA.

R ARIA va furniza fonduri și granturi pentru acoperirea parțială a costurilor de certificare pentru anumite tipuri de dezvoltatori de aplicatii bazate pe IA.

Utilizarea talentelor. O oportunitate pentru România provine din existența unui număr mare de programatori care trebuie susținuți pentru a dezvolta produse și servicii inovative bazate pe IA. Pe această linie, ar fi utile programe specializate pentru susținerea persoanelor fizice.

7. Evaluarea conformității sistemelor de IA

Rezumat. Cadrul pentru reglementarea inteligenței artificiale va avea în centru conceptul de certificare [93]. Această certificare se va realiza în centre de auditare specializate și acreditate de ARIA . Fiind un cadru reglementare de tip NLF, AIA stabilește doar linii directoare, iar operaționalizarea lui se bazează pe implementarea și respectarea de standarde. Prezentăm aici o parte din activitățile de standardizare care vizează aplicații de inteligență artificială. De asemenea, descriem câteva din metodele utilizate (e.g. CRISP-DM, SEMMA) în auditarea sau asigurarea calității sistemelor dezvoltate cu tehnologii din domeniul inteligenței artificiale. Standardele emergente și metodele de auditare a calității prezentate aici pot constitui un punct de plecare pentru cei care doresc organizarea unor astfel de centre de certificare - atât publice cât și private - la nivel național.

7.1 Activități emergente de standardizare a IA

Standardele pentru IA au rolul de (1) a descrie cerințe tehnice uniforme pentru sistemele de IA și (2) a sprijini implementarea cadrelor legale. De asemenea, facilitează accesul pe piață pentru inovațiile IA și oferă furnizorilor de soluții IA un cadru clar pentru dezvoltarea și funcționarea sistemelor IA.

Organizații de standardizare. Organizațiile regionale europene de standardizare, (European Standards Organisations - ESOs), sunt recunoscute oficial de către Comisia Europeană (Regulamentul (UE) nr. 1025/2012) și acționează ca o platformă europeană prin care sunt elaborate standardele europene. Aceastea includ: Comitetul European pentru Standardizare Electrotehnică (CENELEC), Comitetul European pentru Standardizare (CEN) și Institutul European de Standardizare în Telecomunicații (ETSI).

În Uniunea Europeană, numai standardele elaborate de CEN, CENELEC și ETSI sunt recunoscute ca „Standarde europene" (Regulamentul (UE) nr. 1025/2012). ESOs funcționează împreună în interesul armonizării europene, dezvoltând atât standarde solicitate de piață, cât și standarde armonizate în sprijinul legislației europene. Aceste organizații sunt organismele oglindă ale omologilor lor internaționali: Organizația Internațională pentru Standardizare (ISO), Comisia Electrotehnică Internațională (IEC) și Uniunea Internațională a Telecomunicațiilor, sectorul de standardizare a telecomunicațiilor (ITU-T).

■ **Exemplu 7.1 — Organisme de standardizare.** În Germania, Deutsches Institut für Normung (DIN) și Deutsche Kommission für Elektrotechnik, Elektronik, Informationstechnik (DKE) sunt

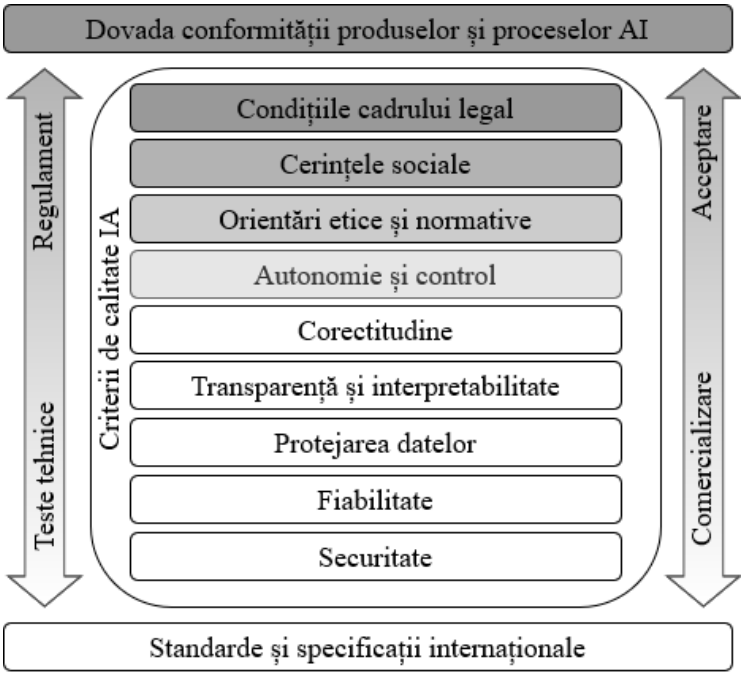

Figura 7.1: Clasificarea criteriilor de calitate IA conform "German Standardization Roadmap on Artificial Intelligence [107]

principalele organisme de standardizare și reprezintă interese naționale la nivelul UE în organizațiile de standardizare precum CEN, CENELEC și ETSI, și la nivel internațional în organizații precum ISO, IEC și UIT [8]. ■

Pentru auditarea sistemelor de IA, se pune întrebarea care sunt standardele de calitate ale IA care necesită testare independentă și ce standarde trebuie elaborate pentru astfel de proceduri de testare. Pentru a aborda această lipsă de standarde, în Germania, de exemplu, a fost prezentată o analiză cuprinzătoare a situației existente și a necesității standardelor și normelor în domeniul inteligenței artificiale sub forma „Normungsroadmap KI" [107]. Cele mai importante dimensiuni de calitate, care ar trebui abordate prin standardizare, sunt prezentate în Fig. 7.1.

Primele standarde care apar în prezent în domeniul IA, în special pentru subiectele fiabilității, robusteții, siguranței și securității sunt enumerate în Tabelul 7.1 Cu toate acestea, este evident că există încă o nevoie considerabilă de dezvoltare în domeniul testării tehnice („testarea produsului"), în special în ceea ce privește validarea și verificarea rețelelor neuronale, argumente de siguranță fiabile pentru sistemele critice de siguranță și instrumente de testare pentru efectuarea acestor teste. Un exemplu important de abordare a acestei nevoi pentru tema conducerii autonome este reprezentat de proiectul german „KI-Absicherung" – Safe AI for Automated Driving. Este administrat de un consorțiu format din instituții de cercetare, producători de automobile, furnizori, organizații de standardizare și autorități publice relevante (precum BSI Germania) și dezvoltă un consens general al industriei cu privire la strategiile de verificare pentru siguranța modulelor bazate pe IA de conducere extrem de automatizată.

O problemă deschisă generală este aceea a compromisurilor (adesea cu mai multe fațete) între caracteristicile dorite ale sistemului, de ex. robustețe, securitate, siguranță și auditabilitate, pe de o parte și caracteristicile modelului IA, algoritmul ML, date și condiții limită, cum ar fi complexitatea modelului, spațiul de activitate, plasticitatea, costul și performanța, pe de altă parte. Aceste compromisuri restricționează scalabilitatea și generalizarea sistemelor actuale de IA.

■ **Exemplu 7.2 — Compromisuri în calitate.** Creșterea complexității modelului poate avea un

Tabela 7.1: Standarde emergente în inteligența artificială

Subiect	**Document**
Fiabilitate și robustețe	ISO/IEC NP 24029: Assessment of robustness of neural networks [53] ITU-T F.AI -DLFE & F.AI-DLPB: metodă de evaluare a cadrului software de învățare profundă și metrică și metode de evaluare pentru benchmark-ul procesorului DNN [36] ETSI DTR INT 008 (TR 103 821): IA în sistemele de testare și testarea modelelor IA, definiții ale metricilor de calitate [1] DIN SPEC 92001-2: Procese ale ciclului de viață IA și cerințe de calitate, Partea 2, robustețe[2] ITU Focus Group pe "Inteligența artificială pentru sănătate" [75] [94]
Siguranță	ISO / CD TR 22100-5: siguranța mașinilor, Partea 5: implicații ale sistemelor embedded IA – ML [56] ISO 26262: vehicule rutiere - siguranță funcțională [100], a se vedea, de asemenea, IEC 61508-1: 2010[3], ISO 21448 [52]) IEEE P2802: standard pentru evaluarea performanței și siguranței dispozitivelor medicale IA - terminologie [48] ISO / IEC AWI TR 5469: IA - siguranță funcțională și sisteme IA[4]
Securitate	ISO / SAE 21434: vehicule rutiere - inginerie de securitate cibernetică[5], s.a. ISO / CD 24089[6], ISO / IEC 23894 [55]) ETSI ISG SAI: Mai multe documente care acoperă declarația problemei, ontologia amenințărilor, testarea securității și strategiile de atenuare pentru sistemele generale de IA (e.g. ETSI GR SAI 005 V1.1.1 (2021-03). 2021 NISTIR 8269: o taxonomie și o terminologie a ML contradictorii [100]

impact negativ asupra interpretabilității și securității. ■

■ **Exemplu 7.3 — Compromisuri în calitate.** Creșterea dimensiunii spațiului de activitate duce la necesitatea unor seturi mai mari de date de testare și validare; ■

■ **Exemplu 7.4 — Compromisuri în calitate.** Întărirea securității duce adesea la performanțe reduse; ■

■ **Exemplu 7.5 — Compromisuri în calitate.** Modelul cutiei albe și accesul la ciclul de viață pentru îmbunătățirea conflictelor de auditabilitate pot intra în conflict cu interesele de proprietate intelectuală; ■

■ **Exemplu 7.6 — Compromisuri în calitate.** Utilizarea seturilor de date externe și a modelelor preantrenate reduce costurile, dar deschide noi vulnerabilități, în special pentru atacurile din spate greu detectabile. ■

Alte documente relevante [108], [38], [74]

Stabilirea priorităților sistemelor de audit IA

Până în prezent nu este disponibil niciun set de criterii și instrumente specifice pentru sistemele de IA. Există două strategii generale care se pot aplica la auditarea aplicațiilor IA:

Crearea condițiilor limită favorabile pentru sarcina dată: o educație adecvată a dezvoltatorilor și utilizatorilor, precum și un schimb suficient de informații între ambele părți permite definirea clară a sarcinii și condițiile limită acceptabile. În combinație cu o analiză ulterioară a riscurilor care ia în considerare încorporarea sistemului IA într-un sistem IT și / sau robotizat mai mare, acesta constituie baza pentru alegeri în cunoștință de cauză în timpul procesului de dezvoltare și a implementării și funcționării sistemului IA. Într-un caz extrem, dezvoltatorul sau utilizatorul ar putea ajunge la concluzia că utilizarea tehnologiei IA trebuie interzisă complet pentru cazul de utilizare specific, de ex. din motive de securitate. În caz contrar, în funcție de caz de utilizare, limitarea spațiului de activitate și limitarea complexității modelului IA pot permite o mai bună auditabilitate și un sistem IA mai sigur și mai sigur [59]. În plus, combinația mai multor măsuri tehnice și organizatorice, precum și, în funcție de considerațiile de proprietate intelectuală, accesul cu cutie albă la modelul și datele sistemului IA pe tot parcursul ciclului de viață în scopuri de evaluare va îmbunătăți cel mai probabil audibilitatea și va contribui la securitate și siguranță.

Investiții în cercetare și dezvoltare: pentru a avansa tehnologiile disponibile și a permite în cele din urmă sisteme sigure de IA, în ciuda condițiilor limită complexe și, prin urmare, pentru a îmbunătăți scalabilitatea și generalizarea. Exemplele includ: a) dezvoltarea unor metrici adecvate pentru toate aspectele relevante pentru securitate și siguranță ale sistemelor de IA. Acestea contribuie la minimizarea impactului compromisurilor, cum ar fi cel dintre performanță și forța de apărare; b) combinația de modele robuste și algoritmi de detecție pentru a respinge intrările potențial rău intenționate, menținând în același timp performanțe ridicate; c) includerea factorilor umani prin ex. modele hibride pentru a îmbunătăți interpretabilitatea; d) generarea eficientă a unui număr mare de atacuri de înaltă calitate, ca bază pentru dezvoltarea unor metode eficiente de apărare, cum ar fi instruirea adversară; e) generarea unei cantități mari de date sintetice realiste de înaltă calitate pentru a contribui la un set de date IID ca bază pentru instruirea sistemelor robuste de IA; f) combinația de simulări realiste cu evaluări din lumea reală și g) utilizarea mai multor sisteme redundante, dar calitativ diferite, de ex. votul majorității sau câștigătorul ia totul.

7.2 Organisme de evaluare a conformității

Cadrul pentru reglementările IA va avea în centru conceptul de certificare IA [93]. Certificarea se realizează de către organisme de evaluare a conformității. Aceste organisme de evaluare a conformității sau centre de auditare sunt acreditate de ARIA .

R Centrele de evaluare a conformității ar putea fi acredidate pe diferite subdomenii ale IA (e.g. învățare automată, sisteme expert, verificare formală etc.) sau pe tipuri de aplicații (e.g. sisteme de recomandare, recunoaștere facială, asistenți virtuali, etc.)

R Centrele de evaluare a conformității ar putea funcționa de exemplu pe lângă Centrele de Cercetare sau Universități, dar și ca entități private (e.g. firme/start-upuri care vor avea ca obiect de activitate certificarea sistemelor bazate pe IA)

Centrele de auditare furnizează servicii de audit tehnic independent a produselor care conțin componente IA. Preconizăm apariția mai multor tipuri de certificate: certificat de etică IA, certificat de siguranță.

■ **Exemplu 7.7 — Malta.** Programul de certificare IA din Malta validează doar că aplicația IA a fost dezvoltată etic și transparent. ■

Certificarea ar putea fi necesară atât la lansarea produsului, dar și la fiecare actualizare software.

ARIA va trebui să obțină gradual expertiza pentru acreditarea organismelor de evaluare a conformității. O posibilă secvență de etape ar putea fi:

1. 2023: Identificarea practicilor de testare și audit (e.g. Care sunt formele de testare și procedurile de auditare pentru a se asigura nediscrimarea? Care sunt măsurile legale și cerințele tehnice pentru garantarea nediscriminarii în sistemele bazate pe IA?) pentru aplicații bazate pe IA în diferite subdomenii (e.g., sisteme autonome, sisteme de recomandare, interacțiune om-IA)
2. 2023: Formularea de recomandări către actorii interesați
3. 2024: Elaborare de standarde

R Regulile pentru testare vor fi publice, astfel încât dezvoltatorii să-și poată verifica în prealabil aplicațiile.

Următoarele aspecte tehnice vizează funcționarea organismelor de evaluare a conformității:

- evaluează riscul de protecție inadecvată a datelor sau algoritmilor.
- evaluează dacă aplicația IA poate introduce efecte anticoncurențiale.
- compară riscurile utilizării aplicatiei IA cu riscurile neutilizării acesteia.
- evaluează necesitatea unor sisteme redundante sau de rezervă, precum și capabilitățile de control și intervenție în caz de eșec a sistemului IA (i.e. abordare orizontală).

■ **Exemplu 7.8 — Evaluarea conformității aplicațiilor IA care interactionează cu copii.** IA va afecta comportamentul copiilor în moduri care nu pot fi înțelese acum. Copiii interacționează deja cu IA prin jucării, asistenți virtuali, jocuri video, sisteme de recomandare. De exemplu, sisteme IA recomandă acum copiilor ce filme să urmarească în continuare, ce cărți să citească, ce prieteni să aibă. Riscurile sunt legate de protecția datelor sau de limitarea dezvoltării personale prin *profiling*, *microtargeting*, *întărirea stereotipurilor*. Părintele are dreptul la explicații, de exemplu "Pe ce criterii ai recomandat acest film copilului meu?" De asemenea, sunt necesare studii comparative pentru "smart toys". Afirmațiile producătorilor precum "îmbunătățesc abilitățile sociale", "îmbunătățesc vocabularul" trebuie probate. Este necesară definirea de metrici specifice

pentru aplicațiile IA care interacționează cu copiii. De asemenea, la evaluarea conformității unei asemenea aplicații IA, ar putea fi oportună introducerea obligativității de a fi inclus și un expert în drepturile copiilor. ■

R Este important să menționăm că certificarea unei aplicații IA nu implică garanții sau validarea performanței, ci implică doar că o organizație terță a validat ca respectiva aplicație IA îndeplinește standardele minime de calitate și/sau bune practici conform grupei de risc în care aceasta se încadrează.

COBIT 2019 Poate fi un punct de pornire pentru auditarea aplicațiilor bazate pe IA deorece furnizează unelte pentru descrierea procesului, ieșirile dorite - pentru toate aplicațiile IT. Pe baza COBIT 2019 se pot compila riscurile și metodele de gestionare a riscurilor legate de aplicațiile IA într-o organizație [51].

■ **Exemplu 7.9 — Criteriu pentru auditare.** Este decizia luată de IA potrivită pe baza datelor de intrare și a cazului de utilizare? [51] ■

Codul de etică al BMW Group pentru inteligența artificială Publicat in 13.10.2020, codul enunță șapte dimensiuni:

1. Implicare umană și supraveghere: monitorizarea umană a deciziilor luate de aplicațiile IA și posibile modalități de anulare a deciziilor luate de algoritmii IA
2. Robusteţe şi siguranţă tehnică: dezvoltarea de aplicații care respectă standardele de siguranță menite să reducă riscul de consecințe sau erori neintenționate
3. Confidenţialitatea şi guvernarea datelor: extinderea măsurilor de confidențialitate și securitate a datelor
4. Transparenţă
5. Diversitate, nediscriminare şi corectitudine: elaborarea aplicațiilor IA ține cont de respectarea demnității umane
6. Grija față de mediu şi societate: dezvoltarea aplicațiilor IA care respectă și promovează bunăstarea clienților, angajaților și a partenerilor.
7. Responsabilitate: identificarea, evaluarea și reducerea riscurilor

■ **Exemplu 7.10 — "Project AI".** A fost lansat în 2018 și are rolul de a sprijini etica și eficiența tehnologiilor IA, fiind centrul de competență al BMW Group pentru analizarea datelor și învățarea automată. Un element principal al Project AI îl reprezintă instrumentul de portofoliu care creează transparență în aplicarea la nivel de companie a tehnologiilor care iau decizii bazate pe date. Portofoliul este numit D3 (decizii bazate pe date – Data Driven Decisions) și acoperă în prezent 400 cazuri de utilizare, dintre care peste 50 sunt disponibile pentru funcționarea regulată. ■

R Ar trebui sprijinită dezvoltarea tehnologiilor și uneltelor care ajută producătorii de IA să dezvolte IA responsabil.

■ **Exemplu 7.11 — Dezvoltarea tehnologiilor pentru IA responsabil.** Verificarea formală este o instrumentație tehnică specifică pentru atenuarea riscurilor relevante la aplicații critice. Verificare modelului (i.e. model checking) unei aplicații IA se realizează prin verificarea unor proprietăți descrise în logici temporale (e.g. Linear Temporal Logic, Computational Tree Logic). ■

7.3 Standarde pentru auditarea sistemelor cu IA

Standardele reprezintă un angajament de calitate către clienți și societate, fiind principala pârghie prin care se va implementa cadrul de reglementare al inteligenței artificiale.

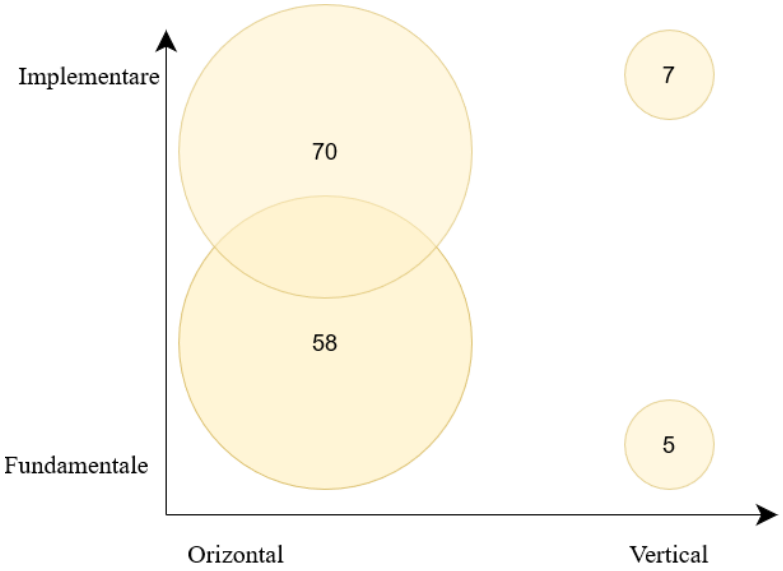

Figura 7.2: Clasificarea standardelor generale legate de IA în funcție de cele două dimensiuni: orizontale/verticale și fundamentale/ implementare.

Nativi și De Nigris au identificat standardele existente și cele în curs de dezvoltare care privesc aplicațiile IA [72]. Metodologia de identificare a acestor standarde a considerat următoarele surse: (i) rezultatele sondajelor existente privind standardizarea IA, de ex. cartea albă a CEN / CELEC Focus Group pe IA, raportul final al proiectului H2020 StandICT.eu (încheiat în 2020), raportul tehnic privind „Standardele pentru guvernarea IA" de la Universitatea din Oxford, studiul eticii IA Grupul de impact pe „Un cadru interdisciplinar pentru operaționalizarea eticii IA", noul Observator European StandICT.eu privind standardele TIC (EUOS); (ii) ublicațiile științifice existente (e.g. Journal of ICT Standardization); (iii) pagini Web ale organizațiilor de standardizare (e.g. ETSI, IEEE, CEN / CENELEC, ISO / IEC JTC1, ITU-T); (iv) foile de parcurs existente privind standardizarea IA: foaia de parcurs ETSI, foaia de parcurs CEN / CENELEC, foaia de parcurs ISO / IEC JTC1, foaia de parcurs germană de standardizare a inteligenței artificiale și foaia de parcurs ITU-T AI; (v) grupuri de lucru, comitete și proiecte care lucrează la standardizarea IA: ISO / IEC JTC1-SC2, EC - CEN CENELEC Focus Group on Artificial Intelligence; Grupul consultativ de experți (EAG) al proiect StandICT.eu (început în 2020 și finanțat de H2020), Comitetul mixt UE-Japonia AI; (vi) Participarea și contribuirea la evenimente specifice care se ocupă de standardizarea TIC șiIA, e.g. seminarul web privind standardizarea IA organizat de DG CNECT (Septembrie 2020), atelierul JRC privind standardizarea (Decembrie 2020), politica JRC privind standardele (Decembrie 2020), reuniunea DGS GROW-CNECT-JRC privind standardele (Ianuarie 2021).

Pasul 1 al metodologiei a fost recunoașterea a aproape 140 de specificații referitoare la IA. Aceste inițiative cuprind atât standarde care abordează în mod direct probleme specifice IA, cât și standarde care sunt tangențial legate de IA, cum ar fi standardele privind tehnologiile abilitate pentru IA, – de exemplu standardele privind Big Data. În pasul al doilea s-au analizat standardele colectate și s-au clasificat în funcție de două dimenisuni (Fig. 7.1).

Standarde identificate de Nativi și De Nigris (72)

Tabela 7.2: Cartografierea la nivel înalt a standardelor semnificative de IA pe cerințele AIA

Cerințe ISO	Date și guvernarea datelor	Sistem de management al riscurilor	Documentația tehnică și păstrarea evidenței	Transparența și furnizarea de informații către utilizatori	Supraveghere umană	Acuratețe, robustețe și securitate cibernetică	Sistem de management al calității
ISO și ISO/IEC JTC1	**ISO/IEC 25024;** ISO/IEC 5259; ISO/IEC 24668;	ISO/IEC 4213; ISO/IEC 25059; ISO/IEC 24029-2	ISO/IEC 5338; ISO/IEC 5469; ISO/IEC 24368; ISO/IEC 24372; ISO/IEC 24668	ISO/IEC 24027; **ISO/IEC 24028;** ISO/IEC 5338; ISO/IEC 24368; ISO/IEC 24372; ISO/IEC 24668; ISO/IEC 4213		ISO/IEC 24027; **ISO/IEC 24028;** ISO/IEC 24029; ISO/IEC 5469	ISO/IEC 23894; ISO/IEC 38507; ISO/IEC 42001; ISO/IEC 25059
IEEE	**ECPAIS Bias;** IEEE P7002; IEEE P7003; IEEE P7004; IEEE P7005; IEEE P7006; IEEE P7009; IEEE P2801; IEEE P2807; IEEE P2863	IEEE P7009; IEEE P2807; IEEE P2846	**ECPAIS Transparență;** IEEE P7000; IEEE P7001; IEEE P7006; IEEE P2801; IEEE P2802; IEEE P2807; IEEE P2863; IEEE P3333.1.3	**ECPAIS Bias; ECPAIS Transparență; ECPAIS Responsabilite;** IEEE P7000; IEEE P7001; IEEE P7003; IEEE P7004; IEEE P7005; IEEE P7007; IEEE P7008; IEEE P7009; IEEE P7011; IEEE P7012; IEEE P7014; IEEE P2863; IEEE P3652.1	**ECPAIS Responsabilite; ECPAIS Transparență;** IEEE P7000; IEEE P7006; **IEEE 7010;** IEEE P7014; IEEE P2863	**ECPAIS Transparență;** IEEE P7007; IEEE P7009; IEEE P7011; IEEE P7012; IEEE P2802; IEEE P2807; IEEE P2846; IEEE P2863; IEEE P3333.1.3	IEEE 2801; IEEE P2863; IEEE P7000
ETSI	DES/eHEALTH-008; **GR CIM 007; GS CIM 009; ENI GS 001;** GR NFV-IFA 041; DGR SAI 002; TR 103 674; TR 103 675; TS 103 327; TS 103 194; TS 103 195.2, SAREF	**GS ARF 003; GR CIM 007; ENI GS 005;** GR NFV-IFA 041; DGS SAI 003; EG 203 341; **TS 103 194; TS 103 195.2;** TR 103 821;	DES/eHEALTH-008; ENI GS 005; DGR SAI 002, **SAREF; GR CIM 007; GS CIM 009**	DES/eHEALTH-008; **GS CIM 009;** DGR SAI 002; **SAREF**	DES/eHEALTH-008; DGR SAI 005	**GS ARF 003; GR CIM 007; ENI GS 001; ENI GR 007;** DGR SAI 001; DGR SAI 002; DGS SAI 003; GR SAI 004; **GS ZSM 002;** TR 103 674; TR 103 675; TS 103 327; **GS 102 181, GS 102 182**	
ITU-T	**ITU-T Y.3170;** ITU-T Y.MecTa-ML; **ITU-T Y.3531; ITU-T Y.3172** ; ITU-T H.CUAV-AIF; ITU-T F.VS-AIMC ; **ITU-T Y.4470;** **Y.Supp.63 și ITU-T Y.4000**	ITU-T Y.qos-ml-arc; **ITU-T Y.3172;** ITU-T H.CUAV-AIF; ITU-T F.VS-AIMC; **ITU-T Y.4470**		**ITU-T Y.4470;**		**ITU-T Y.3170;** ITU-T Y.qos-ml-arc; ITU-T Y.MecTa-ML; **ITU-T Y.3531; ITU-T Y.3172;** ITU-T H.CUAV-AIF; ITU-T F.VS-AIMC; **ITU-T Y.4470**	

Gestiunea datelor	ISO/IEC TS 4213, ISO/IEC 5259-2, ISO/IEC 5259-3, ISO/IEC 5259-4, ISO/IEC 5338, ISO/IEC 5469, ISO/IEC 23894.2, ISO/IEC 24027, ISO/IEC 24029-1, ISO/IEC 24668, ISO/IEC 38507, ISO/IEC 42001,ETSI SAI 002, ETSI SAI 005
Documentație tehnică	ISO/IEC 23894.2, ISO/IEC 24027, ISO/IEC 42001
Înregistrarea operațiunilor	ISO/IEC 23894.2
Transparență și informarea utilizatorilor	ISO/IEC 23894.2, ISO/IEC 24027, ISO/IEC 24028, ISO/IEC 38507, ISO/IEC 42001
Supraveghere umană	ISO/IEC 23894.2, ISO/IEC 38507, ISO/IEC 42001
Acuratețe, robustețe și securitate	ISO/IEC TS 4213, ISO/IEC 5338, ISO/IEC 5469, ISO/IEC 23894.2, ISO/IEC 24029-1, ISO/IEC 24668, ISO/IEC 42001, ETSI SAI 002, ETSI SAI 003, ETSI SAI 005, ETSISAI 006
Gestiunea riscurilor	SO/IEC 5338, ISO/IEC 5469, ISO/IEC 23894.2, ISO/IEC 38507, ISO/IEC 42001
Gestiunea calității	ISO/IEC 5259-3, ISO/IEC 5259-4, ISO/IEC 5338, ISO/IEC 23894.2, ISO/IEC 24029-1, ISO/IEC 38507, ISO/IEC 42001

R O parte din standardele identificate de Nativi et al. [72] sunt în curs de publicare. Nativi et al. a identificat 50 de standarde legate de IA care vor fi publicate până în 2023.

■ **Exemplu 7.12 — Standarde pentru vehicule autonome.** Un astfel de standard este Federal Motor Vehicle Safety Standards (de către NHTSA în SUA) ■

Cadre tehnice standard comune pentru a sprijini implementarea principiilor IA sunt necesare, în special în ceea ce privește problemele de securitate.

- În august 2019, Institutul Național de Standarde și Tehnologie din SUA (NIST) a publicat un raport care subliniază importanța standardelor tehnice IA pentru IA de încredere[7]
- În noiembrie 2020, Germania a lansat foaia de parcurs germană pentru standardizarea inteligenței artificiale[8] ce oferă o imagine de ansamblu a status quo-ului, cerințelor și provocărilor, precum și necesitatea standardizării pe șapte subiecte cheie legate de IA: principii de bază, IA de încredere, calitate, evaluare a conformității și certificare, securitate IT în sistemele IA, automatizare industrială, mobilitate și logistică și IA în medicină. La elaborarea acesteia au fost luate în calcul: Cadrul politic european, inițiativele europene privind IA și strategiile IA din alte țări.
- Organizația Internațională pentru Standardizare (ISO) și Institutul de Ingineri Electrici și Electronici (IEEE) lucreaza la astfel de standarde
- Danemarca, Malta și Suedia, intenționează să stabilească sau să fi stabilit deja programe de certificare IA. Guvernul danez, alături de Confederația industriei daneze, Camera de comerț daneză, SMEdenmark și Consiliul danez al consumatorilor, a creat un sistem independent de etichetare: Sigiliul comun de securitate cibernetică și etică a datelor [22]. Sigiliul este acordat companiilor care îndeplinesc cerințele privind securitatea cibernetică și manipularea responsabilă a datelor legate de IA.

IIA Standard 1210: Proficiency (2) Auditorii interni trebuie să posede cunoștințele, abilități și alte competențe necesare pentru îndeplinirea responsabilităților individuale. Activitatea de audit intern trebuie să posede sau să obțină în mod colectiv cunoștințele, abilitățile și altele competențe necesare pentru

[7] International Telecommunication Union: F.AI-DLFE "Deep Learning Software Framework Evaluation Methodology" (Rev.) Output draft, Virtual meeting, 22 June - 3 July 2020, 2020

[8] DIN. DIN SPEC 92001-2:2020-12 Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 2: Robustheit. 2020.

realizarea acestor responsabilități.

IIA Standard A3: Identificarea elementelor de audit IA bazate pe schema generală de procesare a datelor IA (Figura 7.3):

1. Care este scopul algoritmului de IA?
   - Ce alți algoritmi au fost aplicați pentru a atinge obiective similare și care sunt potențialele problemele asociate acestor algoritmi?
   - Care sunt factorii cheie pentru determinarea rezultatelor algoritmilor?
   - Care sunt posibilele constrângeri de reglementarări în atingerea obiectivului cu și fără aplicarea IA?
2. Obținerea datelor
   - De unde provin datele?
   - Datele de antrenament utilizate în instruirea modelului/mașinii au fost conforme cu cele din sursele de date originale?
   - Datele provin de la o entitate de încredere?
   - Care sunt posibilele neconcordanțe sau alte probleme privind sursele de date, cum ar fi modificările metodologice pentru captarea datelor de-a lungul anilor și problemele de calitate ale sistemelor vechi?
   - Ce criterii au fost utilizate pentru selectarea datelor?
   - Există alte surse de date similare care nu sunt selectate sau folosite pentru antrenarea modelului?
3. Prelucrarea prealabilă a datelor
   - Cum au fost introduse datele lipsă?
   - Care au fost criteriile aplicate pentru eliminarea datelor lipsă?
   - Cum au fost selectate seturile de date de antrenament și testare (stratificare aleatorie, validare încrucișată etc.)?
   - Cum au fost standardizate datele?
4. Modelarea datelor
   - Ce alte tehnici IA au fost luate în considerare, care sunt rezultatele și posibilele motive pentru care acestea nu au fost selectate?
   - Care au fost criteriile și ipotezele considerate pentru analiza algoritmilor?
   - Proiectarea algoritmilor a fost de la zero?
   - Algoritmii folosesc biblioteci de programare și, dacă da, cât de fiabile sunt acestea?
5. Testarea
   - Ce metrici s-au folosit pentru a testa acuratețea modelului?
   - Rezultatele generate de algoritmi au fost sensibile la modificări minore în caracteristicile de modelare?
6. Implementarea
   - Cum a fost instalat modelul în producție și dacă a implicat părți terțe?
   - A existat o revizuire a acurateței după implementarea algoritmilor de către o a treia parte?
   - Algoritmii îndeplinesc obiectivul stabilit la începutul procesului de audit?
7. Monitorizarea rezultatelor
   - Are entitatea structuri, procese și proceduri adecvate pentru a direcționa, gestiona și monitorizează activitățile IA?
   - Ce acțiuni au întreprins părțile responsabile în fiecare etapă a procesului pentru a se asigura că activitățile de IA sunt conforme cu legile și reglementările relevante, în concordanță cu obiectivul organizației și să mențină un nivel adecvat de etică, socială și de responsabilitate?

R Grupul pentru cercetare Tractica Research se așteaptă ca veniturile din IA să crească de la 3.2 miliarde USD în 2016 până la 89.9 miliarde USD în 2025 [51].

Creșterea IA a fost însoțită de timpul tradițional de întârziere între adoptarea timpurie și stabilirea reglementărilor și cadrelor de conformitate. De exemplu, nu există un cadru de audit matur care detaliază subprocesele IA și nici nu există reglementări specifice IA, standarde sau

Figura 7.3: Auditarea sistemelor IA conform schemei generale de procesare a datelor IA

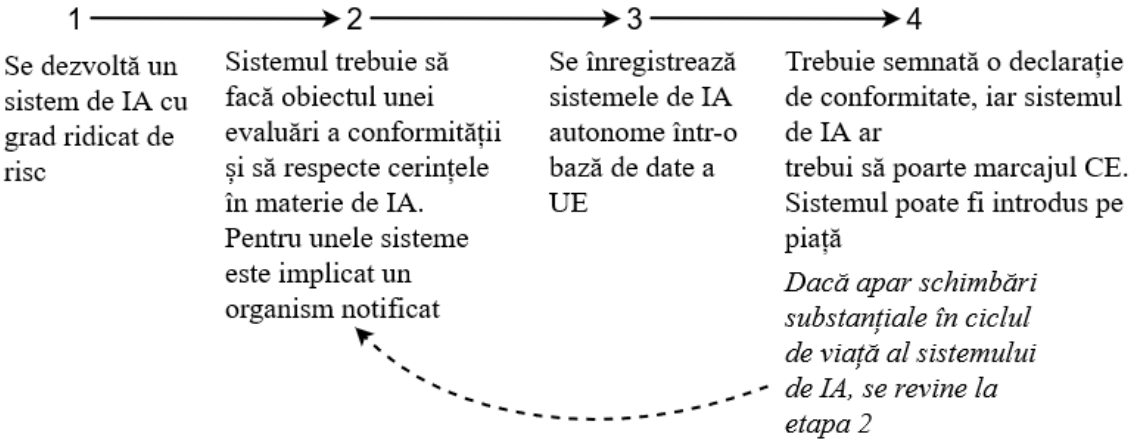

Figura 7.4: Evaluarea aplicațiilor din categoria de risc ridicat

mandate. Un cadru pentru data mining este CRISP-DM, dezvoltat în 2018. Totuși, auditorii au dificultăți deoarece nu există precedente adoptate pe scară largă pentru tratarea cazurilor de utilizare a IA.

În plus față de lipsa unor standarde de audit explicite în jurul IA, există provocări suplimentare care afectează auditul. După cum s-a menționat anterior, definiția IA este dezbătut frecvent și în lumea IT, inclusiv auditorii nu au ajuns la o definiție sau taxonomie comună la care să specifice un set de practici de talie mondială.

Majoritatea companiilor nu se gândesc la modul în care IA poate juca un rol în important în dezvoltarea soluțiilor, așa că este puțin probabil să existe un plan documentat pentru a alinia cazurile de utilizare ale IA în mediul afacerilor sau pentru recunoașterea rentabilității investițiilor în IA.

7.4 Metode de auditare

Fondat în 2016 de Amazon, Facebook, Google, DeepMind, Microsoft și IBM, la care s-au alăturat Apple în 2017 și Baidu în 2018, Partnership on AI este un grup internațional de experți din mediul academic, societatea civilă și industrie. A fost creat pentru a dezvolta cele mai bune practici ale tehnologiilor IA, pentru a ajuta la înțelegerea acestui domeniu de către public și pentru a servi drept platformă de discuții despre IA și impactul acesteia.

În timp ce IA este un instrument puternic pentru combaterea inegalităților, există multe exemple de cazuri în care modelele și seturile de date de învățare automată pot reproduce, sau chiar amplifica, prejudecăți și discriminări. În Statele Unite, instrumentul de recrutare Amazon[9] sau software-ul COMPAS pentru justiție[10] sunt doar ilustrații ale acestui fapt.

Pentru a identifica originea erorilor, a le corecta și a detecta riscurile în timpul fazei de dezvoltare a unei aplicații IA, este necesară monitorizarea și testarea algoritmilor pe tot parcursul ciclului lor de viață, printr-un proces sistematic și documentat.

[9] Amazon scraps secret AI recruiting tool that showed bias against women
[10] Machine Bias

<table>
<tr><th>Scoping</th><th>Mapping</th><th>Artifact Collection</th><th>Testing</th><th>Reflection</th><th>Post-Audit</th></tr>
<tr><td>Definirea scopului în audit IA</td><td>Interesul actorilor relevanți</td><td>Lista de verificare a auditului</td><td>Revizuirea documentației</td><td>Plan de remediere</td><td>Decizii de tip go/no-go</td></tr>
<tr><td>Cerințele produsului IA</td><td>Desfășurarea interviurilor</td><td>Modele de carduri</td><td>Testarea contradictorie</td><td>Fișier cu istoricul proiectărilor</td><td>Atenuări de proiectare</td></tr>
<tr><td>Principiile IA</td><td>Maparea actorilor relevanți</td><td>Fișe tehnice</td><td colspan="2">Diagrama de analizare a riscului în etica IA</td><td>Verificări pe parcurs</td></tr>
<tr><td>Reviziunirea eticii în utilizările IA</td><td>Transcrierea interviurilor</td><td colspan="2"></td><td colspan="2">Redactarea rapoartelor</td></tr>
<tr><td>Evaluarea impactului social</td><td colspan="5">Posibilități de eșec și analiza efectelor (FMEA)</td></tr>
</table>

Figura 7.5: Prezentare generală a cadrului de audit intern. Culoarea gri indică un proces, iar secțiunile colorate reprezintă documente. Documentele în portocaliu sunt produse de auditori, documentele albastre sunt produse de echipele de inginerie și produse și rezultatele verzi sunt dezvoltate în comun.

Complementară noțiunii de explicabilitate, auditailitatea IA descrie posibilitatea de a evalua algoritmi, modele și seturi de date; analiza funcționării, rezultatelor și efectelor sistemelor IA, chiar și pe cele neașteptate. Această noțiune este alcătuită din două părți:

1. *Partea tehnică* constă în măsurarea performanței unui sistem în funcție de mai multe criterii (fiabilitate, precizie a rezultatelor etc.).
2. *Partea etică* constă în reținerea impacturilor sale individuale și colective, precum și verificarea faptului că nu prezintă riscul încălcării anumitor principii, cum ar fi confidențialitatea sau egalitatea. De exemplu, natura nediscriminatorie a unui algoritm de învățare automată va fi testată furnizându-i date de intrare fictive sau profiluri de utilizator.

Deși multe studii subliniază importanța auditurilor externe efectuate de terți după implementarea modelului, ceea ce este interesant în legătură cu metoda propusă de Google AI și Parteneriatul pentru cercetare în IA [85] este abordarea auditului intern în perioada de dezvoltare și pe parcursul întregii faze de proiectare a soluții IA.

Auditurile externe sunt independente și reprezintă mai mult un răspuns la necesitatea de a stabili controale certificate de valoare probatorie. Cu toate acestea, acestea sunt intrinsec limitate de lipsa accesului la procesele și informațiile interne - cum ar fi codul sursă sau datele de instruire - care sunt uneori supuse confidențialității comerciale. De fapt, companii precum Google, care au investit puternic în dezvoltarea sistemelor de IA, sunt reticente în a divulga aceste informații auditorilor externi. Astfel, protecția proprietății intelectuale și a confidențialității în afaceri reprezintă un obstacol major în calea transparenței.

Un audit intern de pre-desfășurare face posibilă intervenția proactivă, nu reactivă, și anticiparea eventualelor erori sau riscuri. Acesta completează auditul extern și îmbunătățește transparența grație producerii, la fiecare etapă a dezvoltării produsului, a unui anumit număr de documente care pot fi consultate de experții externi.

Metoda "Scoping, Mapping, Artifact Collection, Testing, Reflection"'

Inspirat de practicile din alte industrii, cadrul de audit propus de Google AI și Parteneriatul pentru IA poate fi împărțit în cinci etape și se bazează pe seturi de documente numite „artefacte" care sunt produse de auditor și de echipele de dezvoltare. Astfel, a fost dezvoltată metoda SMACTR: Scoping, Mapping, Artifact Collection, Testing și Reflection (Figura 7.5).

Scopul prime etapei (i.e *Scoping*) este de a defini domeniul de aplicare al auditului prin examinarea motivațiilor și a impactului preconizat al sistemului și prin confirmarea principiilor menite să ghideze dezvoltarea acestuia. Acest lucru se face printr-o revizuire etică, care trebuie să includă o varietate de puncte de vedere pentru a maximiza limitarea prejudecăților de „codificare" și printr-o evaluare a impactului social. Scopul acestei etape este de a răspunde la întrebări precum „Cum poate utilizarea sistemului de IA să schimbe viața indivizilor?" și „Care sunt potențialele prejudecăți sociale, economice și culturale?".

Etapa de cartografiere (i.e. *Mapping*) constă în analizarea informațiilor despre diferiții actori și despre ciclul de dezvoltare al produsului. În special, se bazează pe o hartă a diferiților colaboratori, o revizuire a documentației existente și pe rezultatele unui studiu de teren etnografic (realizat prin interviuri cu persoanele cheie ale organizației). Acest studiu trebuie să permită o mai bună înțelegere a modului în care au fost luate anumite decizii, cum ar fi alegerea setului de date sau a arhitecturii modelului, și modul în care acestea vor influența comportamentul sistemului.

În cea de-a treia etapă, denumită etapa de colectare (i.e. *Artifact Collection*), auditorul întocmește un inventar al tuturor documentelor care se presupune că au fost produse în timpul dezvoltării și care sunt necesare pentru a începe auditul. Aceasta include modele de carduri și fișe tehnice, două standarde complementare care vizează îmbunătățirea auditabilității algoritmilor. Cardurile model descriu caracteristicile de performanță ale modelului. Fișele tehnice analizează în special procesul de colectare a datelor și urmăresc să ajute utilizatorul setului de date să ia decizii în cunoștință de cauză.

În timpul etapei de testare (i.e. *Testing*, auditorii interacționează cu sistemul pentru a evalua dacă acesta este conform cu valorile etice ale organizației. Ei pot, de exemplu, inspirându-se din exemple contradictorii (în care sistemul este „păcălit" prin alimentarea cu date de intrare false, astfel încât să se observe rezultatele produse), să trimită profiluri de utilizator false, în special din grupuri, pentru a verifica dacă produce rezultate părtinitoare. Această etapă conține, de asemenea, o diagramă de analiză etică a riscurilor, care ia în considerare combinația probabilității unei defecțiuni (estimată în funcție de apariția anumitor defecțiuni observate în timpul testării) și a severității acesteia (evaluată în etapele anterioare), astfel încât să se definească importanța riscului.

În cele din urmă, etapa de reflecție (i.e. *Reflection*) constă în confruntarea rezultatelor obținute cu așteptările etice predefinite și prezentarea unei analize a riscurilor care descrie principiile care ar putea fi amenințate la implementarea sistemului. Auditorii vor sugera apoi un plan de atenuare a acestor riscuri, obiectivul fiind atingerea unui prag de risc acceptabil predefinit. De exemplu, dacă auditorii descoperă performanțe inegale ale clasificatorului pe subgrupuri, ce nivel de paritate trebuie atins?

Capcana "audit-washing" Auditul intern are anumite limite, care sunt legate în principal de dificultatea pentru auditorul intern de a rămâne independent și obiectiv în timpul executării misiunii lor. Așa cum sistemele de IA nu sunt independente de dezvoltatorii lor, „auditul nu este niciodată izolat de practicile și de persoanele care efectuează auditul" [85]. Pentru a împiedica acest lucru să devină un simplu instrument de marketing, menit să ofere o imagine înșelătoare a responsabilității corporative, auditorii trebuie să fie conștienți de propriile părtiniri și opinii.

R În linii mari, procesul de audit trebuie să fie lent, meticulos și metodic, ceea ce este un contrast puternic cu viteza tipică de dezvoltare a tehnologiilor AI[11]. Poate duce chiar la întreruperea dezvoltării sistemului care este auditat atunci când riscurile depășesc avantajele. Cu toate acestea, este un proces necesar, atât pentru a garanta fiabilitatea, loialitatea și corectitudinea algoritmilor, cât și pentru a permite acceptabilitatea lor socială.

[11] Auditing AI: when algorithms come under scrutiny

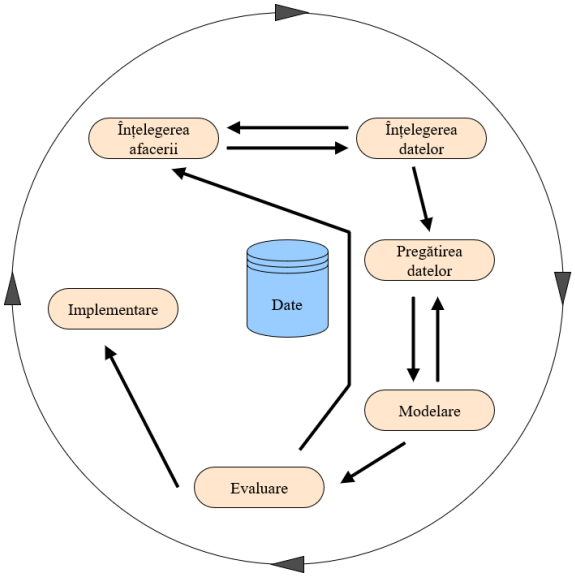

Figura 7.6: Cele 6 faze ale metodei CRISP-DM

Alte documente relevante: [50], [88].

Auditul intern poate ajuta o organizație să evalueze, să înțeleagă și să comunice gradul în care inteligența artificială va avea un efect (negativ sau pozitiv) asupra capacității organizației de a crea valoare pe termen scurt, mediu sau lung. Auditul intern se poate implica prin cel puțin cinci activități critice și distincte legate de inteligența artificială:

1. Pentru toate organizațiile, auditul intern ar trebui să includă IA în evaluarea riscurilor și să ia în considerare dacă trebuie să includă IA în planul său de audit bazat pe risc.
2. Pentru organizațiile care explorează IA, auditul intern ar trebui să fie implicat activ în proiectele de IA încă de la începuturi, oferind sfaturi și informații care să contribuie la implementarea cu succes. Cu toate acestea, pentru a evita percepția sau diminuarea independenței și obiectivității, auditul intern nu ar trebui să dețină și nici să fie responsabil pentru implementarea proceselor, politicilor sau procedurilor de IA.
3. Pentru organizațiile care au implementat un anumit aspect al IA, fie în cadrul operațiunilor sale (cum ar fi un producător care utilizează roboți pe o linie de producție), fie încorporat într-un produs sau serviciu (cum ar fi un comerciant cu amănuntul care personalizează ofertele de produse în funcție de istoricul achizițiilor), auditul intern ar trebui să ofere o asigurare privind gestionarea riscurilor legate de fiabilitatea algoritmilor subiacenți și a datelor pe care se bazează algoritmii.

Metoda CRISP-DM

Cross-industry standard process for data mining [13] (procesul standard transindustrial pentru extragerea datelor), cunoscut sub numele de CRISP-DM, este un model de proces ciclic standard care descrie o abordare structurată a întregului ciclu de viață al datelor, fiind cel mai utilizat model de analiză. Acesta servește nu doar ca o foaie de parcurs a modului de extragere și analiză a datelor, ci și pentru a spori posibilitatea colaborării profesionale. Standardul CRISP-DM este format din șase faze diferite ilustrate în Figura 7.6:

Urmând abordarea CRISP-DM, un nivel de asigurare poate fi obținut printr-o revizuire la nivel înalt, cu mai multă asigurare dacă experții în materie examinează fiecare pas mai în profunzime. Este important de reținut, cadrul CRISP-DM trebuie modificat pentru un audit direcționat spre

diferite domenii, dar totodată să se asigure că pașii corespunzători de lucru sunt documentați.

În comunitatea de învățare automată (machine learning), există un interes considerabil în realizarea de modele de învățare automată interpretabile, în care indivizii pot înțelege că a fost făcută o clasificare dată (de exemplu, acest raport de radiologie arată cancerul, comparativ cu o creștere necanceroasă) în mod intenționat. Acest lucru este important deoarece, în multe domenii, cum ar fi medicina, agenții umani nu vor avea încredere într-un rezultat algoritmic decât dacă se înțelege cum a fost făcută predicția. Există mai multe cadre pentru auditul algoritmilor de învățare automată, cum ar fi cadrul LIME [67] (Local Interpretable Model-Agnostic Explanations) și cadrele FairML [1]. Cu toate acestea, aceste cadre oferă doar o interpretare a ponderilor modelului și nu o înțelegere holistică bazată pe risc a procesului de învățare automată. Aici intră în vigoare abordarea CRISP-DM. Abordările LIME și FairML pot fi utilizate împreună cu cadrul CRISP-DM în etapa de evaluare pentru a asista auditorul în înțelegerea modelului.

Atunci când un model de învățare automată este antrenat și pus în producție, acesta primește date dintr-un set de atribute la un moment dat sau flux de date, în funcție de cazul de utilizare. Pentru acest exemplu, se presupune un model discret cu un singur set de atribute date modelului pe rând. În ambele cazuri, după examinarea parametrilor de intrare, auditorul ar putea obține un pseudo-set de date pentru a fi introdus în algoritm și ar putea examina rezultatele prevăzute pentru caracteristici care ar ajuta la expunerea eventualelor prejudecăți, de exemplu, un model de predicție a împrumutului care discrimină un grup rasial numai prin cod poștal. Prin introducerea datelor într-o gamă de posibilități, se poate obține asigurarea asupra prejudecăților potențiale ale performanței unui model fără a explica pe deplin cum sau de ce algoritmul face o anumită predicție. Chiar și cu un expert în domeniu, o metodă de evaluare a modelului interpretabilă la nivel global nu există în prezent pentru anumite modele (de exemplu, mașini vectoriale de suport și rețele neuronale). Prin evaluarea rezultatului acestei serii de eșantionare (care ar trebui repetată de mai multe ori cu aceleași date de intrare pentru a asigura consistența), precizia practică poate fi determinată în comparație cu precizia matematică utilizată atunci când modelele sunt instruite de oamenii de știință de date (performanța, adică acuratețea modelului și capacitatea acestuia de a îndeplini cerințele afacerii sunt mai puțin complexe de constatat).

1. Înțelegerea afacerii Secțiunea de înțelegere a afacerii ar trebui să fie relativ simplă din perspectiva auditului, dar poate fi o provocare în timpul dezvoltării unui model. Această secțiune abordează înțelegerea cazului de utilizare al afacerii și, cu ajutorul unui expert în domeniu, ce atribute ale cazului de utilizare ar trebui incluse în model, cum ar fi suma venitului, titlul postului, nivelul de educație. În medii sofisticate, când au fost deja utilizate alte tipuri de modele, fie ca software, fie ca modele de decizie mentală, acest pas poate fi puțin mai ușor decât începerea de la zero. Deoarece cadrul CRISP-DM este iterativ, secțiunea de înțelegere a afacerii va fi revizuită adesea în timpul unui proiect de anvergură.

2. Înțelegerea datelor Fără a înțelege natura datelor, nu se poate construi un model precis. Cu toate acestea, acest pas este mai mult decât pare, deoarece majoritatea datelor, pe lângă variabilele categorice, au o scară inerentă, cum ar fi Celsius, Fahrenheit, kilometri, mile etc. Un alt aspect important este locul în care sunt găzduite datele. Diferite depozite de date au considerații diferite, cum ar fi schema dată a unei baze de date relaționale. Fără o înțelegere aprofundată a datelor, nu pot fi realizate modele algoritmice puternice și auditurile ulterioare ale acestora. Auditorul trebuie să fie vigilent în această etapă pentru a înțelege toate variabilele și a se asigura că aceste variabile nu sunt în conflict sau nu introduc prejudecăți. Matricile de corelație și covarianță ar putea fi examinate în această etapă pentru a înțelege modul în care variabilele se corelează și variază ca răspuns unul la altul.

3. Pregătirea datelor Odată ce etapele de înțelegere a cazului de utilizare și colectare a datelor au fost finalizate, este necesară preprocesarea datelor într-o formă utilizabilă pentru modelare. Un avantaj al datelor relaționale este că acestea se pot obține într-o structură modificabilă. În

cazul textului nestructurat, cum ar fi fișierele jurnal și datele casate pe site, etapa de preprocesare poate necesita mult timp. Tehnici precum expresii regulate (regex) pot fi necesare pentru a separa șirurile de text. În majoritatea cazurilor, datele trebuie scalate astfel încât toate caracteristicile sau dimensiunile să aibă aceeași scală. De obicei se utilizează standardizarea scorului z, obținând o medie $\mu = 0$ și o deviație standard $\sigma = 1$.

4. Modelare Modelarea este componenta vitală a învățării automate, însă în majoritatea proiectelor de învățare automată, modelarea este unul dintre pașii mai scurți, cel puțin pentru implementarea inițială. Parametrii pot fi modificați în direcții diferite pentru a rafina performanța unui algoritm. Cu toate acestea, specialiștii folosesc intuiția urmată de tehnici de căutare de tip forță brută pentru a încerca toți hiper-parametrii disponibili (parametri stabiliți înainte de antrenament care nu sunt învățați, într-un anumit interval de valori, obținând rezultatul cel mai bun). În funcție de numărul de hiper-parametri încercați și de complexitatea algoritmului, aceasta poate fi o sarcină foarte intensă din punct de vedere al calculelor. Dacă algoritmii nu au fost reglați, atunci cel mai probabil modelele nu sunt pe deplin optimizate. Acest lucru implică de obicei că nu și-au atins minimele globale, dar lipsa de reglare a modelului nu pune în pericol înțelegerea unui algoritm.

Dezvoltările recente în domeniul învățării automate sunt utilizate din ce în ce mai mult nu numai pentru a regla hiper-parametrii modelelor, ci pentru a selecta algoritmul specific în sine. Atunci când se utilizează învățarea automată, specialistul și auditorul de date trebuie să fie vigilenți și să examineze modelul selectat pentru a se asigura că este îndeplinit gradul de interpretabilitate necesar pentru cazul de utilizare dat. Aceasta înseamnă că afacerea trebuie să explice de ce a fost luată fiecare decizie, ca în cazul companiilor care fac obiectul clauzei „dreptul la explicație" din Regulamentul general de protecție a datelor (GDPR) al Uniunii Europene. În acest context, un model de mașină cu suport vectorial neliniară nu ar fi o alegere acceptabilă. Un beneficiu clar pe care l-a influențat GDPR este acela de a pune mai mult accentul pe interpretabilitatea modelelor în proiectarea algoritmică. În 2016, Conferința internațională de învățare automată (ICML) a început un atelier anual axat pe interpretabilitatea modelelor, numit în mod adecvat „Workshop on Human Interpretability (WHI) in Machine Learning".

Un element extrem de important al listei de verificare a auditului învățării automate ar trebui să examineze dacă datele au fost bifurcate în seturi de instruire și testare. Împărțirea datelor contribuie la prevenirea depășirii modelului, ceea ce înseamnă că algoritmul se potrivește prea strâns cu caracteristicile setului de date individuale, făcând ca acesta să nu se generalizeze bine cu datele noi. În mod tradițional, datele sunt împărțite într-un procentaj 80/20, 80 la sută din date ca date de antrenament și 20 la sută ca date de testare. Cele mai bune practici moderne duc acest lucru un pas mai departe, folosind un proces de validare încrucișată pentru a împărți datele de formare în bucăți mai mici și testarea modelului pe subseturi aleatorii de date. O abordare comună utilizată se numește validare încrucișată K-fold. Aceasta se caracterizează prin împărțirea unui set de date într-un număr K de secțiuni/pliuri în care fiecare pliere este utilizată ca set de testare la un moment dat. De exemplu, scenariul validării încrucișate de 5 ori (K = 5). Aici, setul de date este împărțit în 5 ori. În prima iterație, prima pliere este utilizată pentru a testa modelul, iar restul pentru a antrena modelul. În cea de-a doua iterație, al doilea fold este folosit ca set de testare, în timp ce restul servește ca set de antrenament. Acest proces se repetă până când fiecare pli din cele 5 pliuri a fost folosit ca set de testare.

6. Evaluare Secțiunea de evaluare este, fără îndoială, cea mai importantă secțiune din perspectiva auditului. În acest domeniu al procesului de învățare automată, modelul este validat pentru precizie, iar posibilele erori individuale pot fi evaluate și pentru alte modele. În mod tradițional, modelele sunt evaluate în funcție de precizia predicției și generalizabilitatea față de datele de producție. Cu toate acestea, din perspectiva auditului, o evaluare a rezultatului este o preocupare cheie. Dacă modelul are o precizie de predicție extrem de ridicată (90%) și pare să se generalizeze bine, este posibil să nu îndeplinească obiectivele modelului și/sau să încalce principiile companiei,

Figura 7.7: Metoda de auditare SEMMA

cum ar fi efectuarea accidentală a discriminării rasiale. În plus față de examinarea tuturor pașilor descriși până acum, auditorul ar trebui să creeze un set de date eșantion pentru a-l introduce în algoritm și să evalueze rezultatul pentru a căuta orice efecte neintenționate pe care le poate produce modelul. De exemplu, pentru un model de aprobare a împrumutului, auditorul ar putea crea un set de date cu coduri poștale din cartiere bogate, din clasa de mijloc și din cartiere sărace.

6. Implementare Mai exact, cum și unde este implementat algoritmul în cauză este mai puțin îngrijorător pentru auditor dacă nivelul de serviciu și capacitățile dorite sunt îndeplinite. Cu toate acestea, există un domeniu din care conștientizarea și examinarea auditorilor ar putea oferi valoare (e.g. datoria tehnică). Ori de câte ori un dezvoltator construiește un sistem IA, vor fi luate anumite decizii cu privire la limbajul de programare, APIs, bibliotecile open-source, documentație necesară, câte teste unitare sunt necesare, etc. În esență, datoria tehnică este un factor mai puțin decât ideal, integrat într-un sistem IA. Datoria tehnică nu este inerent rea, ea fiind rezultatul deciziilor luate pentru a realiza proiectele la timp și în limita bugetului. Cu toate acestea, nu este lipsită de consecințe. În învățarea automată, datoria tehnică este mai greu de depistat și remediat decât în proiectele tradiționale de inginerie software datorită aspectului de "învățare". Astfel, accentul este pus pe cascadele de corecție, o varietate insidioasă de datorii tehnice. O cascadă de corecție apare atunci când algoritmul nu produce rezultatul dorit și corecțiile bazate pe reguli sunt aplicate deasupra modelului pentru a corecta deficiențele sale.

Aceste deficiențe pot fi cazuri anterioare sau au apărut din cauza unui model slab sau a unor date de antrenare/validare inadecvate. Problema este că, dacă se aplică prea multe corectii atunci când modelul este antrenat și modificat, devine din ce în ce mai dificil să se constate ce modificări ale modelului vor produce îmbunătățiri, deoarece filtrele sunt în partea de sus a rezultatelor și creează în esență o margine superioară în raport cu capacitatea de învățare a modelului. Datoria tehnică poate fi observată de un om de știință cu experiență și cunoștințe de date care lucrează la model. Cu toate acestea, cunoștințele acumulate dintr-un raport de audit pot consolida necesitatea reorganizării unui model despre care specialiștii știau deja că are datorii tehnice.

Cadrul CRISP-DM a fost introdus pentru a instrui auditorii despre cum să efectueze un audit de învățare automată la nivel înalt. Pentru o aprofundare, va fi necesar un specialist în învățarea automată, dar urmând cadrul dat, auditul învățării automate poate fi accesibil mai multor departamente de audit.

O altă metodă pentru auditarea algorimilor de învățare automat [6] este SEMMA. Metoda *The Sample, Explore, Modify, Model and Access* dezvoltată de Institutul SAS urmează cei 5 pași din Figura 7.7.

R Este necesară pregătirea unor auditori specializați pe inteligență artificială.

7.5 Estimare costuri certificare

Estimare costuri:

1. Costuri cu dezvoltarea asociate îndeplinirii cerințelor: date de antrenare (2,763€), documentare (4,390€), furnizare informații (3,627€), validare de către agentul uman (7,764€), asigurarea

robusteții și acurateței (10,733€), deci 30,000€costuri totale certificare (costuri pe UE 1.6-3.3 miliarde)
2. Costuri pentru certificare: 16,800€–23,000€- ceea ce reprezintă 15% din costul de dezvoltare a unei "unități IA" estimat la 170,000€)

Costurile cu certificarea IA nu ar trebui să depășească un anumit procent din costul de punere în producție (de exemplu, 10 sau 15%), pentru a nu deveni un factor prohibitiv în procesul de inovație și comercializare. Costurile reglementării au fost estimate într-un studiu DG CONNECT realizat pentru Comisia Europeană, la 17% din investiția totală, un procent ridicat și care ar putea duce la scăderea competitivității comerciale a producătorilor sau importatorilor de IA din UE.

R Aceste costuri vor trebui luate în considerare și în cadrul unor finanțări nerambursabile acordate producătorilor de soluții IA, și suportate ca și cheltuieli eligibile.

R La aceste costuri s-ar putea adăuga costuri legate de consultanță pentru obținerea certificării.

R O parte din costuri s-ar putea aplica de fiecare dată când apare o nouă versiune a aplicației bazate pe IA.

R Nu toate aplicațiile de IA au nevoie de certificare. Se estimează că doar 10% din sistemele IA s-ar încadra în grupa de risc ridicat.

R Costurile suplimentare de 10-20% ar putea fi compensate de beneficiile legate de creșterea calității; altfel ar putea exista cheltuieli mai mari cu verificarea, sau costurile accidentelor care sunt evitate.

8. Spații de testare în materie de reglementare

Rezumat. Un instrument important în sprijinirea dezvoltării IA il reprezintă spațiile de testare în materie de reglementare a IA. Acestea reprezintă medii controlate pentru testarea în condiții reale, în România, a sistemelor IA. Este necesar un cadru legal pentru reglementarea funcționalității acestor spații. Viteza cu care se dezvoltă tehnologiile IA ridică în mod recurent provocări legate de identificarea celui mai bun cadru de reglementare. Apare astfel necesitatea de a experimenta eficacitatea unor reglementări, cu scopul de a identifica cele mai bune reguli. Acesta este unul din rolurile spațiilor de testare în materie de reglementare a IA (i.e. "regulatory sandboxes"). Capitolul exemplifică astfel de spații precum și tipurile de clauze de experimentare care apar în aceste medii controlate.

R Spațiile de testare în materie de reglementare a IA oferă un cadru pentru testarea tehnologiilor inovative, dar și pentru a identifica prin experimente reglementările potrivite pentru un produs inovativ.

R Clauzele de experimentare urmăresc furnizarea de "zone de respirație" (i.e "breathing space") pentru dezvoltatorii care au nevoie de a testa noi tehnologii sau modele de afaceri în condiții similare cu mediul real.

Spațiile de experimentare a tehnologiilor și reglementările aferente acestora sunt caracterizate prin [63]:

1. Reprezintă zone de testare, stabilite pentru o perioadă limitată, într-o zonă limitată în care tehnologiile inovatoare sau modele de business inovatoare pot fi testate
2. Experimentele prevăzute sunt controlate prin "clauze de experimentare"
3. Focusul nu e doar pe produsul inovativ, dar și pe identificarea legislației care ar trebui actualizată pentru a acoperi noul produs inovativ.

R Spațiile de experimentare a tehnologiilor și reglementărilor trebuie văzute ca instrumente de sprijin a inovării.

R La nivel național este necesară acumularea de expertiză pentru dezvoltarea de spații de experimentare și reglementare.

R Spațiile de reglementare sunt mecanisme pentru dezvoltarea de reglementări care să țină pasul cu ritmul inovației tehnologice. Ar fi util ca factorii de decizie în politici publice să fie implicați în funcționarea acestor spații.

R Spațiile de testare în materie de reglementare a IA pot fi privite ca instrumente pentru ”politici publice bazate pe dovezi”. Rezultatele experimentelor reprezintă dovezi care vor sta la baza conceperii și justificarii unor eventuale reglementări.

R Este necesar un cadru legal pentru reglementarea funcționalității acestor spații.

R Experimentele desfășurate pot primi finanțare publică. De exemplu, primăria unui oraș finanțează testarea unor autobuze autonome pentru transport public.

R Pentru facilitarea spațiilor de experimentare este util ca mai multe legi și ordonanțe să specifice explicit posibilitatea de utilizare a clauzelor de experimentare“ în contextul acestor spații.

■ **Exemplu 8.1 — Spațiu de reglementare.** Baden-Wurttemberg Autonomous Driving Testbed este un spațiu de reglementare specializat pentru vehicule autonome, care funcționează ca un consorțiu între administrație locală, comunitatea științifică și mediul economic. ■

■ **Exemplu 8.2 — Spații de reglementare.** Vase autonome de navigat pe Dunăre sau pe Canalul Dunare - Marea Neagră. Este nevoie atât de experimentarea tehnologiilor de navigare autonomă, cât și de identificarea reglementărilor potrivite. ■

■ **Exemplu 8.3 — Spații de reglementare.** Experimente pentru identificarea domeniilor din medicină care sunt potrivite pentru telemedicină. În ce măsura pacienții și doctorii răspund pozitiv? Care sunt barierele? Care ar fi modelul de business viabil pentru operatorii aplicațiilor de telemedicină? ■

■ **Exemplu 8.4 — Spații de reglementare: Taxa de trafic în Stockholm.** În 2006 s-a introdus o taxă aplicată aleatoriu, pe o perioadă de 6 luni cu valoare mai mare în intervalul în care traficul e ridicat. Numărul vehiculelor a scăzut cu 20%: jumatate din acești șoferi utilizând transportul în comun, iar jumătate au schimbat intervalul orar. Experimentul acesta a testat și nivelul de acceptare a publicului: înainte de experiment doar 30% din populație era în favoarea taxei; după experiment procentul a crescut la 53% [63]. ■

R Inovația în IA va crea multe situații care nu vor fi clare din perpectiva reglementărilor existente.

Digital Testbed Framework este o inițiativă a administrației din Estonia prin care se oferă acces gratuit la instrumentația tehnică disponibilă la nivel guvernamental (e.g. unelte, tehnologii, date) pentru a favoriza dezvoltarea de produse și servicii inovative. Un avantaj este că testarea acestor aplicații este

asigurată în mediu real. Soluțiile dezvoltate, inclusiv codul, urmează să fie făcute publice și promovate de către guvernul estonian.

R Dacă nu se implementează astfel de spatții de experimentare, produsele inovative vor putea fi testate doar în afara țării.

Recomandare. Ar fi utilă formarea de grupuri de experți din companii, universități și domeniul juridic care să fie asociate acestor spații de testare în materie de reglementare a IA. Expertiza legală plus dialogul cu experții tehnici și dezvoltatorii de soluții IA facilitează conceperea reglementărilor potrivite.

Recomandare. Din considerente de încredere, este util ca spațiile de reglementare să fie inițiate pornind de la rețele sau clustere existente. Între entitățile implicate în diferite experimente trebuie să existe clauze de cooperare.

R Este util ca în experiment să fie invitați și factorii/actorii care pot bloca implementarea produsului. Experimentele vor analiza in ce măsură aceștia sunt afectați și care este cadrul pentru a elimina riscurile ca produsul inovativ să nu fie acceptat de anumiți actori.

Deși aceste spații de testare în materie de experimentare a IA sunt diverse - atât din punctul de vedere al domeniilor de inovare, cât și din punctul de vedere al actorilor implicați, provocările legate de funcționalitatea lor sunt de multe ori aceleași. Pentru a facilita transferul de experiență pe linia organizării unor asemenea spații, în Germania s-a creat Rețeaua Spațiilor de Experimentare [63].

Recomandare Ar fi util un program pilot pentru un hub pentru Standarde IA care să coordoneze și să facilitize implicarea entităților din România în procesele de standardizare a IA.

Un obiectiv al spațiului de reglementare este de a identifica barierele legale care pot bloca introducerea produsului sau serviciului inovativ pe piață. Dacă produsul nu se poate încadra in reglementările existente se identifică ”clauze pentru experimentare“. Acestea pot fi sub forma unei (i) excepții de la prohibiție, (ii) excepție de la un aviz necesar, (iii) excepție pentru furnizarea unei documentații. Aceste tipuri de clauze pot să nu acopere în totalitate anumite situații. În acest caz, se analizează oportunitatea definirii unui nou tip de clauză pentru experimente, sau se recomandă modificarea experimentului pentru a se încadra în tipul de excepții practicat [63].

■ **Exemplu 8.5 — Bariere legale.** Un sistem de telemedicină poate fi blocat de o normă a Colegiului Medicilor conform căruia este interzisă eliberarea de rețete fără a exista interacțiune directă între doctor și pacient. O clauză de experimentare poate introduce o excepție pentru această normă și poate specifica ca aceste eRețete pot fi onorate doar de anumite farmacii (incluse ca parteneri în experiment). ■

R Aprobarea unei clauze pemtru experimentare revine unei autorități de reglementare. O discuție rămâne în ce măsură și pentru ce spețe ARIA poate aproba astfel de clauze, în colaborare sau consultare cu alte autorități.

Chiar dacă clauzele pentru experimentare sunt aprobate, trebuie asigurate măsuri specifice de gestiune a riscurilor. Firmele de asigurare pot oferi reduceri dacă sunt implicate in experiment, iar datele obținute pot fi utilizate de acestea pentru înțelege și riscurile asociate produsului inovativ.

R Spațiile de reglementare în care autoritățile publice sprijină produsul inovativ nu intră sub incidența ”ajutorului de la stat“ , atâta timp cât nu există suport financiar.

Funcționarea spațiului de reglementare se desfășoară sub o autoritate de supraveghere. Aceasta stabilește condiții și comunică condițiile sau evenimentele în care spațiul de reglementare va fi oprit. De asemenea, autoritatea de supraveghere va colecta eventualele plângeri sau probleme apărute în timpul experimentelor.

Facilitarea spațiilor de testare în materie de reglementare a IA face parte dintr-un proces mai larg de asigurare a unui mediu favorabil pentru IA. Configurarea unui astfel de mediu prin tranziția de la cercetare și dezvoltare la comercializarea sau implementarea IA include patru moduri:

1. Furnizarea de medii controlate pentru experimentarea și testarea sistemelor de IA
2. Încurajarea accesului companiilor la finanțare pentru utilizarea acestor instrumente
3. Conectarea companiilor emergente cu oportunități de afaceri prin rețele și platforme de colaborare
4. Furnizarea de consultanță personalizată pentru a sprijini extinderea întreprinderilor. O serie de firme și oferte de servicii precum DataRobot, Amazon Web Services (AWS), Github, Kaggle, Google Tensorflow și servicii conexe contribuie la reducerea barierelor în calea adoptării IA de către firmele mici

Medii controlate pentru experimentarea IA prin testarea în condiții cvasi-reale. Acestea pot oferi o evaluare a impactului tehnologiei IA asupra diferitelor aspecte ale vieții oamenilor (e.g. loc de muncă, educație, mediu). Mediile controlate includ centre de inovare, laboratoare de politici și sandbox-uri de reglementare. Modelele de guvernanță ale co-creației care implică atât guvernele, cât și părțile interesate private, joacă deja un rol cheie în multe strategii naționale de IA [33].

Bibliography

[1] Julius Adebayo. *FairML: ToolBox for diagnosing bias in predictive modeling*. Technical report. Massachusetts Institute of Technology, 2016 (cited on page 113).

[2] Anca Amuza-Conabie et al. *Ghid privind implementarea standardelor internaționale de audit intern*. Technical report. Comitetul de lucru „Audit intern" din cadrul CAFR, 2019. URL: https://www.cafr.ro/wp-content/uploads/2019/12/Ghid-privind-implementarea-Standardelor-internationale-de-audit-intern-2019.pdf (cited on page 107).

[3] France Data Protection Autority. *How can humans keep the upper hand? The ethical matters raised by algorithms and artificial intelligence*. Technical report. France Data Protection Autority, Nov. 2017. URL: %7Bhttps://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199%7D (cited on page 20).

[4] George Antoniu Bara. "Building A Dynamic Corpus Of Fake News Using Commercially Available Machine Translation and NLP Software". In: *2021 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE. 2021, pages 116–121 (cited on page 78).

[5] Chomanski Bartlomiej. "The Missing Ingredient in the Case for Regulating Big Tech". In: *Minds and Machines* (2021), pages 1–19 (cited on page 25).

[6] Jan Roar Beckstrom. "Auditing Machine Learning Algorithms". In: *International Journal of Government Auditing* 48.1 (2021), pages 40–41 (cited on page 115).

[7] Isaac Ben-Israel et al. *Towards Regulation of AI Systems. Global perspectives on the development of a legal framework on Artificial Intelligence (AI) systems based on the Council of Europe's standards on human rights, democracy and the rule of law*. Technical report. CAHAI Secretariat, Dec. 2020. URL: %7Bhttps://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a%7D (cited on page 9).

[8] Christian Berghoff et al. *Towards Auditable AI Systems*. Technical report. Federal Office for Information Security, May 2021 (cited on page 100).

[9] Niels van Berkel et al. “A systematic assessment of national artificial intelligence policies: Perspectives from the Nordics and beyond”. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 2020, pages 1–12 (cited on page 30).

[10] Defense Innovation Board. *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*. Technical report. Defense Innovation Board, Oct. 2019. URL: %7Bhttps://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF%7D (cited on page 20).

[11] Mostafa Bouziane et al. “Team Buster. ai at CheckThat! 2020 Insights and Recommendations to Improve Fact-Checking.” In: *CLEF (Working Notes)*. 2020 (cited on page 83).

[12] Esther Chavannes. “Towards Responsible Autonomy: The Ethics of Robotic and Autonomous Systems in a Military Context”. In: (2019) (cited on page 59).

[13] Andrew Clark. “The machine learning audit CRISP-DM Framework”. In: *ISACA Journal* 1 (2018), pages 42–47 (cited on page 112).

[14] Madeleine de Cock Buning. *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. 2018 (cited on page 62).

[15] European Commission. *Artificial Intelligence for Europe, Communication from the Commission*. Technical report. European Commission, 25 April 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN (cited on page 20).

[16] European Commission. *Coordinated Plan on Artificial Intelligence, COM/2018/795 final*. Technical report. European Commission, July 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52018DC0795&qid=1618308597715 (cited on page 9).

[17] European Commission. *Building trust in human-centric Artificial Intelligence. COM(2019) 168 final*. Technical report. European Commission, Aug. 2019. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52019DC0168&qid=1618308682947 (cited on page 20).

[18] European Commission. *Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services*. Technical report. Official Journal of the European Union, Nov. 2019. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32019R1150&from=EN (cited on page 20).

[19] European Commission. *Report on liability for Artificial Intelligence and other emerging technologies*. Technical report. European Commission, Nov. 2019. URL: %7Bhttps://ec.europa.eu/newsroom/dae/document.cfm?doc_id=63199%7D (cited on page 20).

[20] European Commission. *A European strategy for data, COM(2020) 66 final*. Technical report. European Commission, 19 February 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066&qid=1618308846292 (cited on page 20).

[21] European Commission. *A Union that strives for more*. Technical report. European Commission, 29 January 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0037&qid=1618308771023 (cited on pages 20, 32).

[22] European Commission. *Assessment List for Trustworthy Artificial Intelligence for Self-Assessment*. Technical report. European Commission, 2020 (cited on pages 25, 30, 107).

[23] European Commission. *Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM/2020/825 final*. Technical report. European Commission, Dec. 2020. URL: https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN (cited on page 10).

[24] European Commission. *Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics*. Technical report. European Commission, Feb. 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0064 (cited on page 20).

[25] European Commission. *Report on the safety and liability implications of artificial intelligence, the Internet of Things and robotics. COM(2020) 64 final*. Technical report. European Commission, 19 February 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0064&qid=1618308897061 (cited on page 20).

[26] European Commission. *White Paper On Artificial Intelligence - A European approach to excellence and trust. COM(2020) 65 final*. Technical report. European Commission, 19 February 2020 (cited on pages 20, 32).

[27] European Commission. *Proposal for a Regulation of The European Parliament and of the Council on machinery products*. Technical report. European Commission, Apr. 2021. URL: https://op.europa.eu/en/publication-detail/-/publication/1f0f10ee-a364-11eb-9585-01aa75ed71a1/language-en (cited on page 20).

[28] European Commission. *Proposal for Regulation Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Technical report. European Commission, Apr. 2021. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206 (cited on pages 9, 10, 31, 49, 77, 89, 93).

[29] European Commission. *Regulamentul (UE) 2021/694 al Parlamentului European și al Consiliului din 29 aprilie 2021 de instituire a programului „Europa digitală" și de abrogare a Deciziei (UE) 2015/2240*. Technical report. European Commission, Apr. 2021. URL: https://eur-lex.europa.eu/legal-content/RO/TXT/HTML/?uri=CELEX:32021R0694&qid=1627578734082&from=E (cited on page 49).

[30] European Commission. *The EU's Artificial Intelligence Act - Understand the EU Regulations and which actions to take*. Technical report. European Commission, 2021. URL: https://2021.ai/wp-content/uploads/2021/06/eu-regulations-brochure.pdf (cited on page 10).

[31] Heather A. Conley et al. *The Kremlin Playbook. Understanding Russian Influence in Central and Eastern Europe*. Technical report. Center for Strategic & International Studies, 2020. URL: https://www.csis.org/features/kremlin-playbook-2 (cited on page 64).

[32] Organisation for Economic Cooperation and Development. *Unleashing innovation*. 2019. DOI: https://doi.org/https://doi.org/10.1787/c285121d-en. URL: https://www.oecd-ilibrary.org/content/component/c285121d-en (cited on page 25).

[33] Organisation for Economic Cooperation and Development. *State of Implementation of the OECD AI Principles. Insights from National AI Policies*. Technical report. OECD Digital Economy Papers, 2021 (cited on pages 21, 23, 24, 53, 120).

[34] *Ethics guidelines for trustworthy AI*. Technical report. European Comission, Apr. 2019. URL: %7Bhttps://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai%7D (cited on page 34).

[35] Tatjana Eva. *European framework on ethical aspects of artificial intelligence, robotics and related technologies*. Technical report. he Directorate-General for Parliamentary Research Services (EPRS) of the Secretariat of the European Parliamen, Sept. 2020. URL: %7Bhttps://www.europarl.europa.eu/RegData/etudes/STUD/2020/654179/EPRS_STU(2020)654179_EN.pdf%7D (cited on pages 20, 34).

[36] International Telecommunication Union: F.AI-DLFE. *Deep Learning Software Framework Evaluation Methodology*. Technical report. June 2020 (cited on page 101).

[37] Ed Felten. "Preparing for the future of artificial intelligence". In: *Washington DC: The White House, May* 3 (2016) (cited on pages 22, 92).

[38] David Filip. "An Ontology for Standardising Trustworthy AI". In: () (cited on page 102).

[39] Francesco Cavalli Giorgio Patrini Henry Ajder and Laurence Cullen. *The State of deepfakes*. Technical report. Sensitivity AI, Oct. 2019. URL: https://sensity.%20ai/mapping-the-deepfake-landscape (cited on page 77).

[40] Google. *How Google Fights Disinformation*. Technical report. Google, Feb. 2019. URL: https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf/ (cited on page 74).

[41] Kilian Gross. *AI Act proposal: Article 6 and Annex II*. Technical report. General Secretariat of the Council, DG CNECT, Sept. 2021 (cited on pages 16, 17).

[42] Kilian Gross. *AI Act proposal: Prohibited practices in Art. 5*. Technical report. General Secretariat of the Council, DG CNECT, Sept. 2021 (cited on pages 11–16).

[43] Ad Hoc Expert Group. *Recommendation on the Ethics of Artificial Intelligence (first draft)*. Technical report. UNESCO, July 2020. URL: https://unesdoc.unesco.org/ark:/48223/pf0000373434/ (cited on page 20).

[44] Andrew M Guess, Brendan Nyhan, and Jason Reifler. "Exposure to untrustworthy websites in the 2016 US election". In: *Nature human behaviour* 4.5 (2020), pages 472–480 (cited on page 66).

[45] Jan Hanzelka and Miroslava Pavlıková. "Institutional Responses of European Countries". In: *Challenging Online Propaganda and Disinformation in the 21st Century* (2021), pages 195–224 (cited on pages 63, 66).

[46] Haya R Hasan and Khaled Salah. "Combating deepfake videos using blockchain and smart contracts". In: *IEEE Access* 7 (2019), pages 41596–41606 (cited on page 77).

[47] Amelie Pia Heldt. "Reading between the lines and the numbers: an analysis of the first NetzDG reports". In: *Internet Policy Review* 8.2 (2019) (cited on page 62).

[48] *IEEE. P2802: Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device*. Technical report. 2018 (cited on page 101).

[49] Tencent Research Institute. *Artificial Intelligence A National Strategic Initiative*. Palgrave Macmillan, 2020 (cited on pages 22, 30, 42).

[50] The Institute of Internal Auditors. *Artificial Intelligence - Considerations for the Profession of Internal Auditing*. Technical report. 2017. URL: https://na.theiia.org/periodicals/Public%20Documents/GPI-Artificial-Intelligence.pdf (cited on page 112).

[51] ISACA. *Auditing Artificial Intelligence*. Technical report. 2019. URL: https://ec.europa.eu/futurium/en/system/files/ged/auditing-artificial-intelligence.pdf (cited on pages 104, 108).

[52] ISO. *ISO 21448 Road vehicles - Safety Of The Intended Funktionality*. Technical report. 2019 (cited on page 101).

[53] ISO. *ISO/IEC AWI 24029-2 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*. Technical report. 2021 (cited on page 101).

[54] *ISO 15622 Intelligent transport systems — Adaptive cruise control systems – Performance requirements and test procedures*. Technical report. European Standard, Sept. 2018. URL: https://www.en-standard.eu/iso-15622-intelligent-transport-systems-adaptive-cruise-control-systems-performance-requirements-and-test-procedures (cited on page 47).

[55] *ISO/IEC 23894 Information Technology - Artificial Intelligence - Risk Management*. Technical report. May 2021 (cited on page 101).

[56] *ISO/TR 22100-5:2021 Safety of machinery — Relationship with ISO 12100 – Part 5: Implications of artificial intelligence machine learning*. Technical report. 2021 (cited on page 101).

[57] Anna Jobin, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines". In: *Nature Machine Intelligence* 1.9 (2019), pages 389–399 (cited on page 34).

[58] David MJ Lazer et al. "The science of fake news". In: *Science* 359.6380 (2018), pages 1094–1096 (cited on page 67).

[59] Mathias Lechner et al. "Neural circuit policies enabling auditable autonomy". In: *Nature Machine Intelligence* 2.10 (2020), pages 642–652 (cited on page 102).

[60] Stefan Leijnen et al. "An agile framework for trustworthy AI." In: *NeHuAI@ ECAI*. 2020, pages 75–78 (cited on page 34).

[61] Chris Lewis. *Irresistible Apps: Motivational design patterns for apps, games, and web-based communities*. Apress, 2014 (cited on page 12).

[62] Radu Magdin et al. ""Disinformation campaigns in the European Union: Lessons learned from the 2019 European Elections and 2020 Covid-19 infodemic in Romania"". In: *Romanian Journal of European Affairs* 20.2 (2020), pages 49–61 (cited on page 64).

[63] *Making space for innovation. The Handbook for regulatory sandboxes*. Federal Ministry for Economic Affairs and Energy (BMWi), July 2019 (cited on pages 117–119).

[64] Bertin Martens et al. *The digital transformation of news media and the rise of disinformation and fake news*. Technical report. European Commission, Joint Research Centre, Apr. 2018. URL: https://ec.europa.eu/jrc/sites/default/files/jrc111529.pdf (cited on pages 61, 66, 67).

[65] Sean McGregor. "Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database". In: *CoRR* abs/2011.08512 (2020). arXiv: 2011.08512. URL: https://arxiv.org/abs/2011.08512 (cited on page 91).

[66] Hannah Miller and Richard Stirlin. *Government Artificial Intelligence Readiness Index2019*. Technical report. Canada's International Development Research Center, Apr. 2019. URL: https://africa.ai4d.ai/wp-content/uploads/2019/05/ai-gov-readiness-report_v08.pdf (cited on page 51).

[67] Saumitra Mishra, Bob Sturm, and Simon Dixon. "Local Interpretable Model-Agnostic Explanations for Music Content Analysis." In: *ISMIR*. 2017, pages 537–543 (cited on page 113).

[68] Lisa Monaco et al. *Survey of Global Artificial Intelligence Regulation: An Evolving and Varied Landscape*. Technical report. O'Melveny & Myers LLP, 2020 (cited on page 29).

[69] Preslav Nakov et al. "Overview of the CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News". In: *Proceedings of the 12th International Conference of the CLEF Association: Information Access Evaluation Meets Multiliguality, Multimodality, and Visualization*. CLEF '2021. Bucharest, Romania (online), 2021 (cited on page 85).

[70] Preslav Nakov et al. "The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news". In: *European Conference on Information Retrieval*. Springer. 2021, pages 639–649 (cited on page 85).

[71] United Nations. *Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. Technical report. United Nations, 2015 (cited on page 54).

[72] Stefano. Nativi and Sarah De Nigris. "AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework, EUR 30772 EN". In: (2021). URL: %7Bhttps://publications.jrc.ec.europa.eu/repository/handle/JRC125952%7D (cited on pages 19, 105, 107).

[73] Ministerul Apărării Naționale. *Carta albă a apărării, București*. Technical report. Ministerul Apărării Naționale, 2021. URL: http://legislatie.just.ro/Public/DetaliiDocumentAfis/242221 (cited on page 55).

[74] Luis Oala et al. "ML4h auditing: From paper to practice". In: *Machine Learning for Health*. PMLR. 2020, pages 280–317 (cited on page 102).

[75] L. Oala et al. "Ml4h auditing: From paper to practice". In: *Data and artificial intelligence assessment methods (DAISAM),* ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting I 2020. 2020 (cited on page 101).

[76] James Pamment, Howard Nothhaft, and Alicia Fjallhed. *Countering Information Influence Activities: A Handbook for Communicators*. MSB, 2018. ISBN: 978-91-7383-867-2. URL: https://www.msb.se/RibData/Filer/pdf/28698.pdf (cited on page 64).

[77] James Pamment, Howard Nothhaft, and Alicia Fjällhed. "Countering information influence activities: A handbook for communicators". In: (2018). URL: https://www.msb.se/RibData/Filer/pdf/28698.pdf (cited on pages 64, 65).

[78] Argyri Panezi. "Liability Rules For AI-Facilitated Wrongs: An Ecosystem Approach To Manage Risk And Uncertainty". In: *AI And The Law (forthcoming volume, Pablo Garcia Mexia & Francisco Perez Bes, eds.)* (2021) (cited on page 30).

[79] Kostantinos Papadamou et al. "Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children". In: *Proceedings of the international AAAI conference on web and social media*. Volume 14. 2020, pages 522–533 (cited on page 74).

[80] P J Phillips et al. "Four Principles of Explainable Artificial Intelligence (Draft)". In: (2020) (cited on page 26).

[81] Sandra Planes-Satorra and Caroline Paunov. "The digital innovation policy landscape in 2019". In: 71 (2019). DOI: https://doi.org/https://doi.org/10.1787/6171f649-en. URL: https://www.oecd-ilibrary.org/content/paper/6171f649-en (cited on page 25).

[82] Eugen-Florian Popescu. "An Integrated Analysis on the Influence of Online Education on Students' Learning Process". In: *International Journal of Education and Research* 9.6 (2021), pages 61–82 (cited on page 53).

[83] United States. Executive Office of the President. "Artificial intelligence, automation, and the economy". In: (2016) (cited on page 22).

[84] *Protection from Online Falsehoods and Manipulation Act (POFMA)*. Technical report. POFMA Office, 2019. URL: https://www.pofmaoffice.gov.sg/regulations/protection-from-online-falsehoods-and-manipulation-act (cited on page 63).

[85] Inioluwa Deborah Raji et al. "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pages 33–44 (cited on pages 110, 111).

[86] A Renda. "Europe: toward a policy framework for trustworthy AI". In: *The Oxford Handbook of Ethics of AI* (2020), pages 649–666 (cited on page 33).

[87] Repatriot. *România în era inteligenței artificiale. Recomandări pentru dezvoltarea și adaptarea tehnologiei IA la nivel de țară*. Technical report. Repatriot, 2020 (cited on page 23).

[88] Asociatia Auditorilor Interni din Romania. *Ghid privind implementarea Standardelor international de audit intern 2019*. Technical report. 2019. URL: https://www.cafr.ro/wp-content/uploads/2019/12/Ghid-privind-implementarea-Standardelor-internationale-de-audit-intern-2019.pdf (cited on page 112).

[89] Niklas H. Rossbach. *An analysis of Germany,s NetzDG law*. URL: https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf (cited on page 62).

[90] Andrew Roth. "Combating deepfake videos using blockchain and smart contracts". In: *The Guardian* (2021) (cited on page 77).

[91] Victoria L Rubin. "News verification suite: Towards system design to supplement reporters' and editors' judgements". In: *Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI*. 2017 (cited on page 83).

[92] S Samoili et al. *Defining artificial intelligence : towards an operational definition and taxonomy of artificial intelligence*. Technical report. Joint Research Centre (European Commission, 27 February 2020. URL: https://op.europa.eu/en/publication-detail/-/publication/6cc0f1b6-59dd-11ea-8b81-01aa75ed71a1/language-en (cited on page 10).

[93] Matthew Scherer. "Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies". In: *Harv. JL & Tech.* 29 (2015), page 353 (cited on pages 90, 99, 103).

[94] E. Schörverth et al. *FG-AI4H Open Code Initiative - Evaluation and Reporting Package, ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H)*. Technical report. 2021 (cited on page 101).

[95] Estonian Foreign Intelligence Service. *"International Security and Estonia 2021*. Technical report. 2021. URL: https://www.valisluureamet.ee/pdf/raport/2021-ENG.pdf (cited on page 65).

[96] Shaden Shaar et al. "Overview of the CLEF-2021 CheckThat! Lab Task 1 on Check-Worthiness Estimation in Tweets and Political Debates". In: *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*. CLEF '2021. Bucharest, Romania (online), 2021 (cited on page 85).

[97] Shaden Shaar et al. “Overview of the CLEF-2021 CheckThat! Lab Task 2 on Detecting Previously Fact-Checked Claims in Tweets and Political Debates”. In: *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*. CLEF ’2021. Bucharest, Romania (online), 2021 (cited on page 85).

[98] Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. “Overview of the CLEF-2021 CheckThat! Lab Task 3 on Fake News Detection”. In: *Working Notes of CLEF 2021—Conference and Labs of the Evaluation Forum*. CLEF ’2021. Bucharest, Romania (online), 2021 (cited on page 85).

[99] Singularex. *“The Black Market for Social Media Manipulation*. Technical report. Riga: NATO Strategic Communications Centre of Excellence, 2019 (cited on page 65).

[100] Elham Tabassi et al. “A taxonomy and terminology of adversarial machine learning”. In: *NIST IR* (2019) (cited on page 101).

[101] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. “Particular object retrieval with integral max-pooling of CNN activations”. In: *arXiv preprint arXiv:1511.05879* (2015) (cited on page 69).

[102] Department of Transportation. *Federal Automated Vehicles Policy*. Technical report. European Parliamentary Research Service, 2016. URL: https://www.transportation.gov/AV/federal-automated-vehicles-policy-september-2016 (cited on page 47).

[103] Heidi Tworek and Paddy Leerssen. *Psychological Defence: Vital for Sweden’s Defence Capability*. 2019 (cited on page 64).

[104] Scientific Foresight Unit. *Tackling deepfakes in European policy*. Technical report. European Parliamentary Research Service, July 2021 (cited on pages 76, 77).

[105] Sabine Verheyen. *Report on artificial intelligence in education, culture and the audiovisual sector*. Technical report. Committee on Culture and Education, Apr. 2021 (cited on page 53).

[106] Jean-Baptiste Jeangene Vilmer. *“Hybrid CoE Research Report 2: Effective state practices against disinformation: Four country case studies*. Technical report. he European Centre of Excellence for Countering Hybrid Threa, July 2021. URL: https://www.hybridcoe.fi/wp-content/uploads/2021/07/20210709_Hybrid_CoE_Research_Report_2_Effective_state_practices_against_disinformation_WEB.pdf (cited on pages 64, 65).

[107] W Wahlster and C Winterhalter. “German standardization roadmap on artificial intelligence”. In: *DIN/DKE, Berlin/Frankfurt* (2020) (cited on page 100).

[108] Philip Matthias Winter et al. “Trusted Artificial Intelligence: Towards Certification of Machine Learning Applications”. In: *arXiv preprint arXiv:2103.16910* (2021) (cited on page 102).

[109] Bernd W Wirtz, Jan C Weyerer, and Carolin Geyer. “Artificial intelligence and the public sector—applications and challenges”. In: *International Journal of Public Administration* 42.7 (2019), pages 596–615 (cited on page 51).

[110] Herbert Zech. “Liability for AI: public policy considerations”. In: *ERA Forum*. Volume 22. 1. Springer. 2021, pages 147–158 (cited on page 30).

Glosar

AIA Artificial Intelligence Act (Actul pentru Inteligență Artificială). 5, 12–14, 16, 18, 19, 23, 29, 45, 99, 106

AIID AI Incident Database. 91

Algorithmic Impact Assessment este un instrument de evaluare a riscului unui sistem de decizie automat. Instrumentul este public și este bazat pe un chestionar care evaluează factori precum proiectarea sistemului, algoritmii utilizați, tipul de decizie, calitatea datelor utilizate. 92

Algorithmic Impact Assessment Tool Instrument de evaluare a impactului algoritmilor în administrația publică utilizat de guvernul canadian. 91

Bias de confirmare atrofierea abilităților persoanelor care interactionează și își bazează deciziile pe aplicații IA . 28

Clickbait metodă de exploatare a curiozității oamenilor prin titluri și imagini atractive care au legătura redusă cu articolul în scopul de chiar daca nu au legătura cu articolul cu scopul de genera vizitatori și bani din publicitate. 67, 68

CRISP-DM Cross-industry standard process for data mining. 99

Dark Patterns șabloane intunecate care manipulează utilizatorul prin elemente de interfață. 12

Deepfake reprezintă manipularea conținutului audio sau video pentru a atribui unei persoane afirmații sau acțiuni pe care aceasta nu le-a facut, folosindu-se tehnici din inteligența artificială (e.g. rețele adversariale generative, autoencodere). 76

Dreptul la explicație dreptul personanelor de a primi o explicație a unei decizii care a fost luată pe baza recomandării unui algoritm sau aplicații bazate pe IA. 90, 91

Dreptul la informare dreptul persoanelor de a fi informate dacă interacționează cu un sistem bazat pe IA (e.g. chatbot). 90

DSA Digital Services Act (Actul pentru Servicii Digitale). 13

Federated learning metodă distribuită de învățare computațională în care fiecare dispozitiv îmbunătățește un model preantrenat pe baza propriilor exemple de antrenamente, noul model rezultat fiind facut public pentru utilizare de către alte dispozitive. 47

GAN Rețele Adversariale Generative - arhitecturi pentru învățarea automată care pe baza unor imagini de intrare generează imagini sintetice asemenătoare cu un nivel similar de calitate. 93

German Observatory for Artificial Intelligence instituție pentru reglementare și monitorizare a IA în Germania. 94, 95

Human Rights Impact Assessments este un instrument pentru evaluarea implicațiilor sistemelor bazate pe IA asupra drepturilor omului. 92

NLF New Legislative Framework (Noul Cadru Legislativ). 17, 18, 99

Principiul "once-only" Principiu de guvernanță în Estonia prin care administrației îi este interzis să ceară de la un cetățean o informație de două ori. 96

RBI Remote Biometric Identification (identificare biometrică la distanță). 15, 18

RPA Robotic Process Automation. 46, 48

SDG Sustainable Development Goals, cele 17 obiective conform agendei pentru dezvoltare sustenabilă pentru 2030 adoptată de către UN în 2015. 45

SEMMA Sample, Explore, Modify, Model and Access process. 99, 115

Spațiu de testare în materie de reglementare a IA reprezintă zone unde reglementările sunt limitate și favorabile pentru testarea aplicațiilor bazate pe IA. 91

The Institute for AI and Ethical ML https://ethical.institute. 40

UCPD Unfair Commercial Practice Directive (Directiva cu privire la practicile comerciale incorecte). 13

Verificarea formală instrumentație din inteligența artificială prin care proprietățile unei aplicații critice sunt verificate cu ajutorul unor instrumente precum logicile temporale. 104

XAI Explainable AI urmărește ca soluția propusă de sistemele cu IA să fie explicată și înțeleasă de agentul uman. 26, 43