GRUPUL INDEPENDENT

DE EXPERȚI LA NIVEL ÎNALT

PRIVIND INTELIGENȚA ARTIFICIALĂ

INSTITUIT DE COMISIA EUROPEANĂ ÎN IUNIE 2018

ORIENTĂRI ÎN MATERIE DE ETICĂ PENTRU O INTELIGENȚĂ ARTIFICIALĂ (IA) FIABILĂ

ORIENTĂRI ÎN MATERIE DE ETICĂ PENTRU O INTELIGENȚĂ ARTIFICIALĂ (IA) FIABILĂ

Grupul de experți la nivel înalt privind inteligența artificială

Prezentul document a fost redactat de Grupul de experți la nivel înalt privind IA (AI HLEG). Membrii Al HLEG menționați în prezentul document sprijină cadrul global pentru o inteligență artificială (IA) fiabilă prevăzut în prezentele orientări, deși nu sunt neapărat de acord cu fiecare afirmație din document. .

Lista de evaluare pentru o IA fiabilă, prezentată în capitolul III din prezentul document, va fi inclusă într-o etapă pilot, în cadrul căreia părțile interesate vor colecta feedback practic. O versiune revizuită a listei de evaluare, ținând seama de feedbackul obținut în etapa pilot, va fi prezentată Comisiei Europene la începutul anului 2020.

Grupul de experți la nivel înalt privind inteligența artificială (Al HLEG) este un grup independent de experți, care a fost instituit de Comisia Europeană în iunie 2018.

Contact Nathalie Smuha - coordonatorul AI HLEG
E-mail CNECT-HLG-AI@ec.europa.eu

Comisia Europeană
B-1049 Brussels

Document publicat la X aprilie 2019.

Un prim proiect al prezentului document a fost publicat la 18 decembrie 2018 și a făcut obiectul unei consultări deschise care a generat feedback de la peste 500 de contribuitori. Dorim să le mulțumim în mod explicit și călduros tuturor celor care au contribuit cu feedbackul lor privind primul proiect al documentului, care a fost luat în considerare la elaborarea prezentei versiuni revizuite.

Mai multe informații cu privire la Grupul de experți la nivel înalt privind inteligența artificială sunt disponibile online (https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence).

CUPRINS

REZUMAT **2**

A. INTRODUCERE **5**

B. UN CADRU PENTRU O IA FIABILĂ **7**

I. Capitolul I: Fundamentele IA fiabile **10**

1. Drepturile fundamentale ca drepturi morale și legale 11
2. De la drepturile fundamentale la principiile etice 11

II. Capitolul II: Asigurarea IA fiabile **16**

1. Cerințele pentru o IA fiabilă 16
2. Metodele tehnice și metodele fără caracter tehnic pentru realizarea unei IA fiabile 24

III. Capitolul III: evaluarea IA fiabile **28**

C. EXEMPLE DE OPORTUNITĂȚI ȘI DE PREOCUPĂRI MAJORE GENERATE DE IA **40**

D. CONCLUZIE **44**

GLOSAR **46**

REZUMAT

(1) Orientările au scopul de a promova o IA fiabilă. IA fiabilă are **trei componente**, care ar trebui să fie îndeplinite pe toată durata ciclului de viață al sistemului: (a) IA ar trebui să fie **legală**, respectând toate legile și reglementările aplicabile; (b) ar trebui să fie **etică**, asigurând respectarea principiilor și a valorilor etice și (c) ar trebui să fie **solidă**, atât din perspectivă tehnică, cât și socială, deoarece, chiar dacă au intenții bune, sistemele IA pot provoca daune neintenționate. Fiecare componentă în sine este necesară, dar nu suficientă pentru a dezvolta o IA fiabilă. În mod ideal, toate cele trei componente funcționează în armonie și se suprapun. În cazul în care, în practică, apar tensiuni între aceste componente, societatea ar trebui să depună eforturi pentru a le alinia.

(2) Prezentele orientări stabilesc un **cadru pentru dezvoltarea unei inteligențe artificiale fiabile**. Cadrul nu tratează în mod explicit prima componentă a IA fiabile (IA legală)[1]. În schimb, acesta are scopul de a oferi îndrumări privind promovarea și asigurarea unei IA etice și solide (a doua și a treia componentă). Adresându-se tuturor părților interesate, prezentele orientări își propun să nu se limiteze la o listă de principii etice, oferind îndrumare cu privire la modul în care acestor principii li se poate oferi o dimensiune operațională în cadrul unor sisteme socio-tehnice. Aceste orientări sunt structurate pe trei niveluri de abstractizare, de la cele mai abstracte, prevăzute în capitolul I, la cele mai concrete, prevăzute în capitolul III, încheind cu exemple de oportunități și preocupări majore generate de sistemele IA.

I. Pe baza unei abordări întemeiate pe drepturile fundamentale, capitolul I identifică **principiile etice** și valorile corelate acestora care trebuie să fie respectate la dezvoltarea, implementarea și utilizarea sistemelor IA.

Orientări esențiale rezultate din capitolul I:

- ✓ Dezvoltarea, implementarea și utilizarea sistemelor IA astfel încât să se respecte următoarele principii etice: *respectarea autonomiei oamenilor, prevenirea daunelor, echitatea și explicabilitatea.* Recunoașterea și abordarea tensiunilor potențiale dintre aceste principii.
- ✓ Acordarea unei atenții deosebite situațiilor care implică mai multe grupuri vulnerabile, cum ar fi copiii, persoanele cu handicap și alte grupuri care au fost defavorizate la nivel istoric sau care sunt expuse riscului de excluziune, precum și situațiilor caracterizate de asimetrii ale puterii sau ale informațiilor, cum ar fi cele dintre angajatori și lucrători sau dintre întreprinderi și consumatori[2].
- ✓ Recunoașterea și luarea în considerare a faptului că, deși aduc beneficii substanțiale persoanelor fizice și societății, sisteme IA pot totodată prezenta anumite riscuri și pot avea un impact negativ, inclusiv un impact care poate fi dificil de anticipat, de identificat sau de măsurat (de exemplu, asupra democrației, statului de drept și justiției distributive sau chiar asupra minții umane). Adoptarea unor măsuri adecvate pentru abordarea acestor riscuri atunci când este necesar și în mod proporțional cu magnitudinea riscului.

II. Având la bază capitolului I, capitolul II oferă orientări cu privire la modul în care se poate realiza IA fiabilă, enumerând **șapte cerințe** pe care sistemele IA ar trebui să le îndeplinească. Pentru punerea lor în aplicare pot fi utilizate atât metode tehnice, cât și metode fără caracter tehnic.

---

[1] Toate afirmațiile normative din prezentul document au scopul de a reflecta orientările în vederea dezvoltării celei de a doua și a treia componente a IA fiabile (IA etică și IA solidă). Prin urmare, aceste afirmații nu sunt menite să furnizeze consultanță juridică sau să ofere orientări cu privire la respectarea legilor aplicabile, deși se recunoaște faptul că multe dintre aceste afirmații sunt reflectate deja într-o anumită măsură în legile existente. În această privință, a se vedea punctul 21 și următoarele.

[2] A se vedea articolele 24-27 din Carta drepturilor fundamentale a Uniunii Europene (Carta UE), care consacră drepturile copilului și ale persoanelor în vârstă, integrarea persoanelor cu handicap și drepturile lucrătorilor. A se vedea, de asemenea, articolul 38 care consacră protecția consumatorilor.

Orientări esențiale rezultate din capitolul II:

- ✓ Asigurarea faptului că dezvoltarea, implementarea și utilizarea sistemelor IA îndeplinesc cerințele pentru IA fiabilă: (1) factorul uman și supravegherea umană, (2) soliditatea tehnică și siguranța, (3) viața privată și guvernanța datelor, (4) transparența, (5) diversitatea, nediscriminarea și echitatea, (6) bunăstarea socială și ecologică și (7) responsabilitatea.
- ✓ Analizarea metodelor tehnice și fără caracter tehnic pentru a se asigura punerea în aplicare a acestor cerințe.
- ✓ Stimularea cercetării și a inovării pentru a contribui la evaluarea sistemelor IA și a sprijini îndeplinirea cerințelor; diseminarea rezultatelor și a întrebărilor deschise pentru publicul larg și formarea în mod sistematic a unei noi generații de experți în etica IA.
- ✓ Comunicarea, într-un mod clar și proactiv, de informații părților interesate cu privire la capacitățile și limitările sistemului IA, pentru a le permite să aibă așteptări realiste, precum și informații 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 a posibilității de auditare a sistemelor IA, în special în contexte sau situații critice.
- ✓ Implicarea părților interesate pe toată durata ciclului de viață al sistemului IA. Stimularea formării și a educației, astfel încât toate părțile interesate să fie informate în ceea ce privește IA fiabilă și să fie formate în acest sens.
- ✓ Conștientizarea faptului că ar putea exista tensiuni fundamentale între diferitele principii și cerințe. Identificarea, evaluarea, documentarea și comunicarea în mod continuu a compromisurilor și a soluțiilor găsite în acest sens.

III. Capitolul III prezintă o Lista concretă și neexhaustivă de evaluare pentru o IA fiabilă, care are scopul de a oferi o dimensiune operațională cerințelor stabilite în capitolul II. Această **listă de evaluare** va trebui să fie adaptată utilizării specifice a sistemului IA[3].

Orientări esențiale rezultate din capitolul III:

- ✓ Adoptarea unei evaluări a IA fiabile atunci când se dezvoltă, se implementează ori se utilizează sistemele IA și adaptarea acesteia în funcție de utilizarea specifică a sistemului.
- ✓ Luarea în considerare a faptului că această listă de evaluare nu va fi niciodată exhaustivă. Asigurarea faptului că IA fiabilă nu înseamnă bifarea unor căsuțe, ci identificarea în mod continuu și punerea în aplicare a cerințelor, evaluarea soluțiilor și asigurarea unor rezultate îmbunătățite pe parcursul întregului ciclu de viață al sistemului IA și implicarea părților interesate în acest proces.

(3) O secțiune finală a documentului are scopul de a concretiza unele aspecte abordate în întregul cadru, oferind exemple de oportunități avantajoase care ar trebui să fie valorificate și de preocupări majore generate de sistemele IA care ar trebui să fie luate în considerare cu atenție.

(4) Deși prezentele orientări au scopul de a oferi îndrumare pentru aplicațiile IA în general, având la bază o platformă orizontală pentru a dezvolta IA fiabilă, situații diferite generează provocări diferite. Prin urmare, ar trebui să se analizeze dacă, pe lângă acest cadru orizontal, este necesară o abordare sectorială, având în vedere caracterul specific al sistemelor IA în funcție de context.

(5) Prezentele orientări nu sunt destinate să înlocuiască nicio formă de elaborare a politicilor sau de reglementare

[3] În conformitate cu sfera cadrului stabilit la punctul 2, această listă de evaluare nu oferă consultanță cu privire la asigurarea conformității legale (IA legală), ci se limitează la a oferi orientări cu privire la îndeplinirea celei de a doua și a treia componente ale IA fiabile (IA etică și IA solidă).

actuală sau viitoare ori să descurajeze inițierea unor astfel de politici ori reglementări. Orientările ar trebui să fie considerate un document evolutiv, care trebuie să fie revizuit și actualizat de-a lungul timpului pentru a se asigura relevanța sa continuă, deoarece atât tehnologia, cât și mediile noastre sociale și cunoștințele noastre evoluează. Prin urmare, prezentul document este conceput ca un punct de plecare pentru discuția privind „O inteligență artificială fiabilă pentru Europa".[4]Dincolo de frontierele Europei, aceste orientări vizează, de asemenea, stimularea cercetării, a reflecției și a discuțiilor privind un cadru etic pentru sistemele IA la nivel mondial.

[4] Acest ideal este menit să se aplice sistemelor IA dezvoltate, implementate și utilizate în statele membre ale UE, precum și sistemelor dezvoltate și produse în afara UE, dar implementate și utilizate în UE. Când se face trimitere la „Europa" în prezentul document, acest termen cuprinde statele membre ale UE. Totuși, prezentele orientări se doresc, de asemenea, să fie relevante în afara UE. În această privință, se poate remarca, de asemenea, că atât Norvegia, cât și Elveția fac parte din Planul coordonat privind IA, stabilit și publicat în decembrie 2018 de către Comisie și statele membre.

A. INTRODUCERE

(6) În comunicările sale din 25 aprilie 2018 și din 7 decembrie 2018, Comisia Europeană (Comisia) și-a prezentat viziunea pentru inteligență artificială (IA), care sprijină „o IA etică, sigură și de vârf realizată în Europa”[5]. Viziunea Comisiei are la bază trei piloni: (i) consolidarea investițiilor publice și private în IA pentru a stimula adoptarea acesteia; (ii) pregătirea pentru schimbări socioeconomice și (iii) asigurarea unui cadru etic și juridic adecvat pentru a consolida valorile europene.

(7) Pentru a sprijini punerea în aplicare a acestei viziuni, Comisia a instituit Grupul de experți la nivel înalt privind inteligența artificială (AI HLEG), un grup independent mandatat să elaboreze două documente: (1) Orientările în materie de etică pentru IA și (2) Recomandările în materie de politică și de investiții.

(8) Prezentul document conține Orientările în materie de etică pentru IA, care au fost revizuite în urma deliberărilor în cadrul grupului nostru, în lumina feedbackului primit ca urmare a consultării publice privind proiectul publicat la 18 decembrie 2018. Acesta se bazează în continuare pe activitatea Grupului european pentru etică în domeniul științei și al noilor tehnologii[6] și se inspiră din alte eforturi similare[7].

(9) În ultimele luni, 52 de membri s-au reunit, au discutat și au interacționat, în spiritul devizei europene: Uniți în diversitate. Considerăm că IA are potențialul de a transforma societatea în mod semnificativ. IA nu reprezintă un scop în sine, ci mai degrabă un mijloc promițător de a spori prosperitatea umană, consolidând astfel bunăstarea individuală și socială și binele comun și aducând progrese și inovație. În special, sistemele IA pot contribui la facilitarea realizării obiectivelor de dezvoltare durabilă ale ONU, cum ar fi promovarea echilibrului de gen și combaterea schimbărilor climatice, raționalizarea utilizării resurselor naturale de către oameni, consolidarea sănătății noastre, a mobilității și a proceselor de producție, precum și sprijinirea modului în care monitorizăm progresele în raport cu indicatorii de durabilitate și coeziune socială.

(10) Pentru a face acest lucru, sistemele IA[8] trebuie să fie **centrate pe om**, bazându-se pe un angajament față de utilizarea acestora în serviciul umanității și al binelui comun, cu obiectivul de a îmbunătăți bunăstarea și libertatea oamenilor. Deși oferă oportunități importante, sistemele IA comportă, de asemenea, anumite riscuri care trebuie să fie gestionate în mod adecvat și proporțional. În prezent, dispunem de o șansă importantă să conturăm dezvoltarea acestora. Dorim să asigurăm faptul că putem avea încredere în mediile socio-tehnice în care acestea sunt încorporate și dorim ca producătorii de sisteme IA să obțină un avantaj concurențial prin încorporarea unei IA fiabile în produsele și serviciile lor. Acest lucru implică încercarea de a **maximiza beneficiile oferite de sistemele IA**, **prevenind și diminuând**, în același timp, **riscurile acestora**.

(11) În contextul unor schimbări tehnologice rapide, considerăm că este esențial ca încrederea să rămână fundamentul societăților, al comunităților, al economiilor și al dezvoltării durabile. Prin urmare, identificăm **IA fiabilă ca fiind ambiția noastră fundamentală**, deoarece oamenii și comunitățile vor putea să aibă încredere în dezvoltarea tehnologiei și în aplicațiile acesteia atunci când va exista un cadru clar și cuprinzător pentru asigurarea fiabilității acesteia.

(12) Considerăm că aceasta este calea pe care Europa ar trebui să o urmeze pentru a se poziționa drept cămin și lider pentru o tehnologie etică și de vârf. Prin intermediul IA fiabile noi, cetățenii europeni, vom încerca să

---

[5] COM(2018)237 și COM(2018)795. Trebuie remarcat faptul că sintagma „realizată în Europa” este utilizată în textul comunicării Comisiei. Totuși, domeniul de aplicare al prezentelor orientări cuprinde nu doar sistemele IA realizate în Europa, ci și sistemele dezvoltate în afara Europei și implementate sau utilizate în Europa. Prin urmare, în cadrul prezentului document, intenționăm să promovăm IA fiabilă „pentru” Europa.

[6] Grupul european pentru etică în domeniul științei și al noilor tehnologii (EGE) este un grup consultativ al Comisiei.

[7] A se vedea secțiunea 3.3 din COM(2018)237.

[8] Glosarul de la finalul prezentului document oferă o definiție a sistemelor IA în scopul prezentului document. Această definiție este explicată în mod detaliat într-un document specific elaborat de AI HLEG care însoțește prezentele orientări, intitulat „O definiție a inteligenței artificiale (IA): principalele capacități și discipline științifice”.

valorificăm beneficiile acesteia într-o manieră aliniată la valorile noastre fundamentale ale respectării drepturilor omului, democrației și statului de drept.

IA fiabilă

(13) Fiabilitatea este o condiție prealabilă pentru ca oamenii și societățile să dezvolte, să implementeze și să utilizeze sistemele IA. Dacă sistemele IA și oamenii din spatele lor nu sunt cu adevărat demni de încredere, pot apărea consecințe nedorite și ar putea fi îngreunată adoptarea tehnologiei, ceea ce ar împiedica realizarea beneficiilor sociale și economice potențial importante pe care le aduc sistemele IA. Pentru a ajuta Europa să realizeze aceste beneficii, viziunea noastră este de a utiliza etica drept pilon principal pentru a asigura și a dezvolta IA fiabilă.

(14) Încrederea în dezvoltarea, implementarea și utilizarea sistemelor IA vizează nu numai proprietățile intrinsece ale tehnologiei, ci și calitățile sistemelor socio-tehnice care implică aplicații IA[9]. În mod similar problemelor de încredere (sau de pierdere a încrederii) în aviație, energia nucleară sau siguranța alimentară, nu numai componentele sistemului IA, ci și sistemul în ansamblu pot genera încrederea sau o pot pierde. Prin urmare, eforturile pentru realizarea IA fiabile se referă nu numai la fiabilitatea sistemului IA în sine, ci necesită și o abordare holistică și sistemică, care să cuprindă fiabilitatea tuturor actorilor și a tuturor proceselor care fac parte din contextul socio-tehnic al sistemului pe toată durata ciclului său de viață.

(15) IA fiabilă are **trei componente**, care ar trebui să fie îndeplinite pe toată durata ciclului de viață al sistemului:

1. IA ar trebui să fie **legală**, asigurând respectarea tuturor legilor și a reglementărilor aplicabile;
2. IA ar trebui să fie **etică**, asigurând respectarea principiilor și a valorilor etice; și
3. IA ar trebui să fie **solidă**, atât din perspectivă tehnică, cât și socială, deoarece, chiar dacă au intenții bune, sistemele IA pot provoca daune neintenționate.

(16) Fiecare dintre aceste trei componente este necesară, dar nu este suficientă în sine pentru a obține o IA fiabilă[10]. În mod ideal, toate cele trei funcționează în armonie și se suprapun. Totuși, în practică, pot exista tensiuni între aceste elemente (de exemplu, uneori domeniul de aplicare și conținutul legislației existente ar putea să nu fie în concordanță cu normele etice). Este responsabilitatea noastră individuală și colectivă, ca societate, să depunem eforturi pentru a asigura faptul că toate cele trei componente contribuie la asigurarea IA fiabile[11].

(17) O abordare fiabilă este esențială pentru a permite o „competitivitate responsabilă", asigurând fundamentul în baza căruia toți cei afectați de sistemele IA să poată avea încredere că proiectarea, dezvoltarea și utilizarea acestora sunt legale, etice și solide. Prezentele orientări au scopul de a stimula inovarea responsabilă și durabilă în domeniul IA în Europa. Orientările urmăresc să transforme etica într-un pilon principal pentru dezvoltarea unei abordări unice a IA, cu scopul de a capacita și a proteja atât prosperitatea umană individuală, cât și binele comun al societății, și de a aduce beneficii acestora. Considerăm că acest lucru va permite Europei să se poziționeze ca lider global în ceea ce privește IA de ultimă generație, demnă de încrederea noastră individuală și colectivă. Doar prin asigurarea fiabilității, cetățenii europeni vor valorifica pe deplin beneficiile sistemelor IA, având siguranța faptului că există măsuri de protecție împotriva riscurilor potențiale.

(18) La fel cum utilizarea sistemelor IA nu se oprește la frontierele naționale, nici impactul lor nu face acest lucru. Prin urmare, sunt necesare soluții globale pentru oportunitățile și provocările globale pe care le generează IA. Așadar, încurajăm toate părțile interesate să depună eforturi pentru realizarea unui cadru global pentru IA fiabilă, ajungând la un consens internațional, promovând și susținând, în același timp, abordarea noastră

[9] Aceste sisteme includ oameni, actori statali, corporații, infrastructură, software, protocoale, standarde, guvernanță, legislații existente, mecanisme de supraveghere, structuri de stimulente, proceduri de audit, raportare privind bunele practici și altele.

[10] Acest lucru nu exclude faptul că s-ar putea impune condiții suplimentare.

[11] Acest lucru înseamnă, de asemenea, că poate fi necesar ca organismele legislative sau factorii de decizie să revizuiască adecvarea legislației existente în cazul în care aceasta ar putea să nu fie în concordanță cu principiile etice.

bazată pe drepturile fundamentale.

Publicul și domeniul de aplicare

(19) Prezentele orientări se adresează tuturor părților interesate din domeniul IA care proiectează, dezvoltă, implementează, pun în aplicare, utilizează sau sunt afectate de IA, inclusiv, dar fără a se limita la societăți, organizații, cercetători, servicii publice, agenții guvernamentale, instituții, organizații ale societății civile, persoane fizice, lucrători și consumatori. Părțile interesate care s-au angajat la realizarea IA fiabile pot opta în mod voluntar să utilizeze prezentele orientări ca metodă de operaționalizare a angajamentului lor, în special utilizând lista de evaluare practică din capitolul III în cadrul proceselor lor de dezvoltare și implementare a sistemelor IA. De asemenea, această listă de evaluare poate completa procesele de evaluare existente și, prin urmare, poate fi integrată în aceste procese.

(20) Orientările au scopul de a oferi îndrumare pentru aplicațiile IA în general, având la bază o platformă orizontală pentru a realiza IA fiabilă. Cu toate acestea, **situații diferite generează provocări diferite**. Sistemele IA de recomandări muzicale nu generează aceleași preocupări etice ca sistemele IA care propun tratamente medicale esențiale. De asemenea, sistemele IA utilizate în contextul relațiilor business-to-consumer, business-to-business, angajator-angajat și public-cetățean sau, în termeni mai generali, în diferite sectoare sau situații de utilizare generează oportunități și provocări diferite. Prin urmare, având în vedere caracterul specific al sistemelor IA în funcție de context, se recunoaște că punerea în aplicare a prezentelor orientări trebuie să fie adaptată la aplicația IA în cauză. În plus, ar trebui explorată necesitatea unei abordări sectoriale suplimentare, pentru a completa cadrul orizontal mai general propus în prezentul document.

Pentru a înțelege mai bine modul în care pot fi puse în aplicare aceste orientări la nivel orizontal, precum și aspectele care necesită o abordare sectorială, invităm toate părțile interesate să testeze Lista de evaluare pentru o IA fiabilă (capitolul III), care operaționalizează acest cadru, și să ne ofere feedback. Pe baza feedbackului obținut prin această etapă pilot, vom revizui lista de evaluare din prezentele orientări până la începutul anului 2020. Etapa pilot va fi lansată până în vara anului 2019 și va dura până la sfârșitul anului. Toate părțile interesate vor putea participa, manifestându-și interesul prin intermediul Alianței europene în domeniul inteligenței artificiale.

B. UN CADRU PENTRU O IA FIABILĂ

(21) Prezentele orientări formulează un cadru pentru realizarea IA fiabile pe baza drepturilor fundamentale consacrate în Carta drepturilor fundamentale a Uniunii Europene (Carta UE) și în legislația internațională privind drepturile omului. În continuare, analizăm pe scurt cele trei componente ale IA fiabile.

IA legală

(22) Sistemele IA nu funcționează într-o lume fără legi. O serie de norme cu caracter juridic obligatoriu la nivel european, național și internațional se aplică deja sau sunt relevante pentru dezvoltarea, implementarea și utilizarea sistemelor IA în prezent. Sursele juridice relevante includ, dar nu se limitează la, dreptul primar al UE (Tratatele Uniunii Europene și Carta drepturilor fundamentale a UE), dreptul derivat al UE (cum ar fi Regulamentul general privind protecția datelor, directivele anti-discriminare, Directiva privind echipamentele tehnice, Directiva privind răspunderea pentru produsele cu defect, Regulamentul privind libera circulație a datelor fără caracter personal, legislația pentru consumatori și directivele privind securitatea și sănătatea în muncă), dar și tratatele ONU privind drepturile omului și convențiile Consiliului Europei (cum ar fi Convenția europeană a drepturilor omului) și numeroase legislații ale statelor membre ale UE. Pe lângă normele aplicabile la nivel orizontal, există diverse norme specifice domeniului, care se aplică anumitor aplicații IA (cum ar fi, de exemplu, Regulamentul privind dispozitivele medicale din sectorul sănătății).

(23) Legislația prevede atât obligații pozitive, cât și negative, adică nu ar trebui să fie interpretată doar cu trimitere

la ceea ce *nu* se poate face, ci cu trimitere la ceea ce *ar trebui* să se facă. Legislația nu doar interzice anumite acțiuni, ci și permite alte acțiuni. În această privință, se poate observa că Carta UE conține articole privind „libertatea de a desfășura o activitate comercială” și „libertatea artelor și științelor”, împreună cu articole care abordează domenii cu care suntem mai familiarizați atunci când încercăm să asigurăm fiabilitatea IA, cum ar fi, de exemplu, protecția datelor cu caracter personal și nediscriminarea.

(24) Orientările nu tratează în mod explicit prima componentă a IA fiabile (IA legală), ci oferă, în schimb, îndrumare privind stimularea și asigurarea celei de a doua și a treia componente (IA etică și IA solidă). Deși ultimele două sunt adesea reflectate deja într-o anumită măsură în legislațiile existente, realizarea lor deplină poate să nu se limiteze la obligațiile legale existente.

(25) Nicio dispoziție din prezentul document nu se va interpreta ca furnizând consultanță juridică sau orientări cu privire la modul în care se poate realiza conformitatea cu normele juridice și cerințele existente aplicabile. Nicio dispoziție din prezentul document nu creează drepturi legale și nici nu impune obligații legale față de terțe părți. Cu toate acestea, amintim faptul că este obligația oricărei persoane fizice sau juridice să respecte legile - indiferent dacă acestea sunt aplicabile în prezent sau vor fi adoptate în viitor în funcție de dezvoltarea IA. Prezentele orientări pornesc de la prezumția că **toate drepturile și obligațiile legale care se aplică proceselor și activităților implicate în dezvoltarea, implementarea și utilizarea IA rămân obligatorii și trebuie să fie respectate în mod corespunzător**.

IA etică

(26) Pentru dezvoltarea IA fiabile, este necesar nu numai să se respecte legea, care constituie una dintre cele trei componente ale sale. Legile nu sunt întotdeauna în pas cu evoluțiile tehnologice, uneori acestea pot fi depășite de normele etice sau, pur și simplu, pot să nu fie potrivite pentru a aborda anumite probleme. Prin urmare, pentru ca sistemele IA să fie fiabile, acestea ar trebui totodată să fie etice, asigurând alinierea la normele etice.

IA solidă

(27) Chiar dacă se asigură un scop etic, persoanele fizice și societatea trebuie, de asemenea, să aibă încredere că sistemele IA nu vor provoca daune neintenționate. Aceste sisteme ar trebui să funcționeze în condiții de siguranță și securitate și într-un mod fiabil și ar trebui să se prevadă dispozitive de securitate pentru a preveni orice impact negativ neintenționat. Prin urmare, este important să se asigure faptul că sistemele IA sunt solide. Acest lucru este necesar atât din perspectivă tehnică (asigurarea soliditățiitehnice a sistemului dacă este cazul într-un context dat, cum ar fi domeniul de aplicare sau faza ciclului de viață), cât și din perspectivă socială (ținând seama, în mod corespunzător, de contextul și mediul în care funcționează sistemul). Prin urmare, IA etică și IA solidă sunt strâns legate și se completează reciproc. Principiile prezentate în capitolul I și cerințele care rezultă din acestea în capitolul II abordează ambele componente.

Cadrul

(28) Orientările din prezentul document sunt structurate pe trei niveluri de abstractizare, de la cele mai abstracte, prevăzute în capitolul I, la cele mai concrete, prevăzute în capitolul III:

(I) Fundamentele IA fiabile. Capitolul I stabilește fundamentele IA fiabile, prezentând abordarea acesteia bazată pe drepturile fundamentale[12]. Acesta identifică și descrie principiile etice care trebuie să fie respectate pentru a asigura IA etică și IA solidă.

(II) Asigurarea IA fiabile. Capitolul II traduce aceste principii etice în șapte cerințe pe care sistemele IA ar

[12] Drepturile fundamentale stau atât la baza legislației internaționale privind drepturile omului, cât și a dreptului Uniunii privind drepturile omului și susțin drepturile aplicabile legal, garantate de Tratatele UE și de Carta drepturilor fundamentale a UE. Având caracter juridic obligatoriu, respectarea drepturilor fundamentale se încadrează, prin urmare, în prima componentă a IA fiabile, „IA legală”. Totuși, drepturile fundamentale pot fi înțelese și ca reflectând drepturi morale speciale ale tuturor persoanelor fizice, care rezultă în virtutea umanității lor, indiferent de caracterul lor juridic obligatoriu. În acest sens, drepturile fundamentale fac parte, de asemenea, din a doua componentă a IA fiabile, „IA etică”.

trebui să le pună în aplicare și să le îndeplinească pe toată durata ciclului lor de viață. În plus, acesta oferă atât metode tehnice, cât și metode fără caracter tehnic, care pot fi utilizate pentru punerea lor în aplicare.

(III) Evaluarea IA fiabile. Practicienii în domeniul IA așteaptă orientări concrete. Prin urmare, capitolul III stabilește o listă de evaluare preliminară și neexhaustivă pentru o IA fiabilă, pentru a operaționaliza cerințele din capitolul II. Această evaluare ar trebui să fie adaptată la aplicarea specifică a sistemului.

(29) Secțiunea finală a documentului prezintă oportunitățile benefice și preocupările majore pe care le generează sistemele IA, care ar trebui să fie luate în considerare și cu privire la care am dori să stimulăm dezbaterile ulterioare.

(30) Structura orientărilor este ilustrată în *figura 1* de mai jos

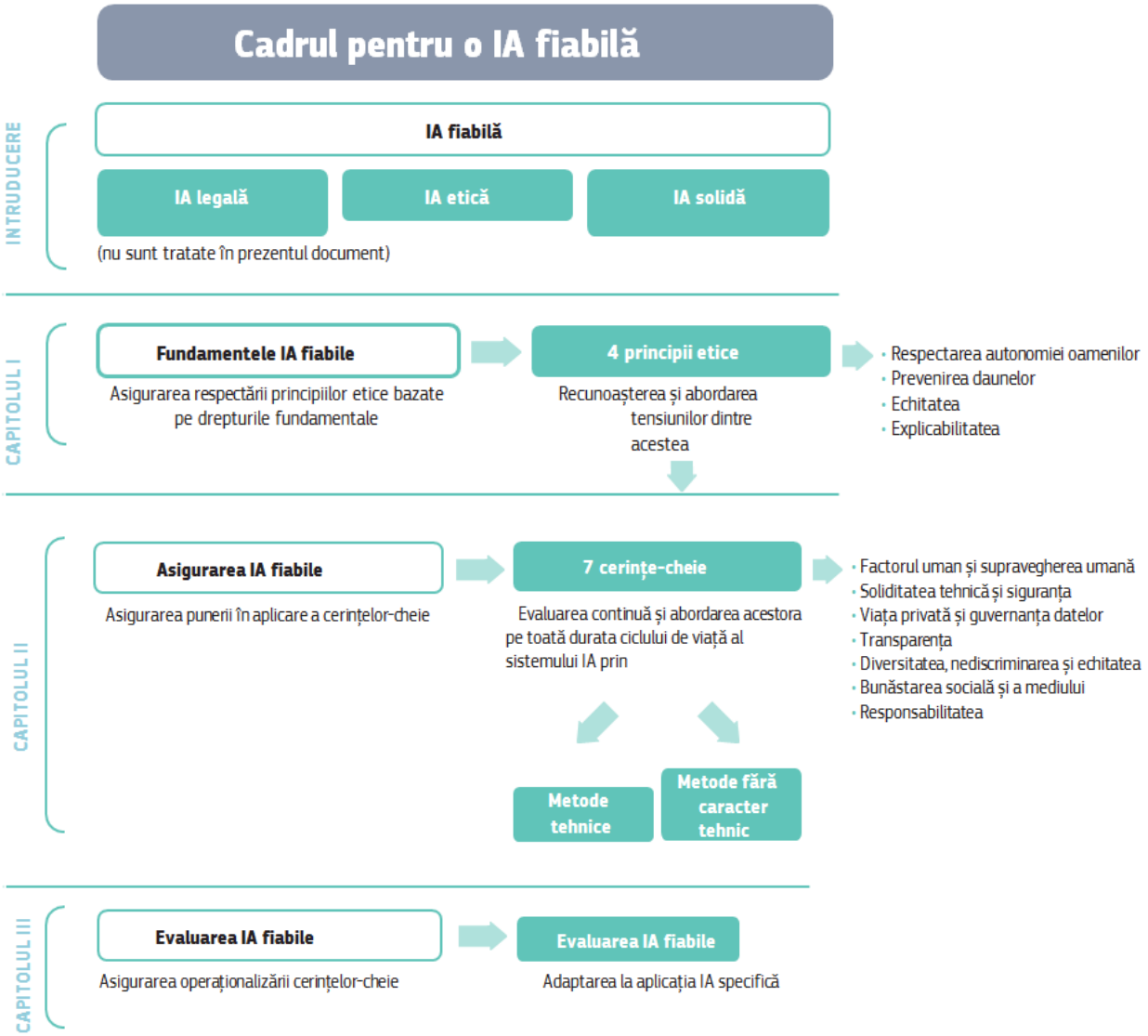

Figura 1: Orientările drept cadru pentru o IA fiabilă

I. Capitolul I: Fundamentele IA fiabile

(31) Acest capitol stabilește fundamentele IA fiabile, întemeiate pe drepturile fundamentale și reflectate de cele patru principii etice care ar trebui să fie respectate pentru a asigura IA etică și IA solidă. Acest capitol se inspiră în mare măsură din etică.

(32) Etica IA este un subdomeniu al eticii aplicate, care se axează pe problemele etice generate de dezvoltarea, implementarea și utilizarea IA. Principala sa preocupare este aceea de a identifica modul în care IA poate să evolueze sau să genereze preocupări pentru viața bună a persoanelor fizice, atât în ceea ce privește calitatea vieții, cât și autonomia oamenilor și libertatea necesare pentru o societate democratică.

(33) Reflecția etică asupra tehnologiei IA poate servi unor scopuri multiple. În primul rând, aceasta poate stimula reflecția cu privire la necesitatea de a proteja persoanele fizice și grupurile la nivelul de bază. În al doilea rând, aceasta poate stimula noi tipuri de inovare care urmăresc să promoveze valorile etice, cum ar fi cele care contribuie la realizarea obiectivelor de dezvoltare durabilă ale ONU[13], încorporate în mod ferm în viitoarea Agendă 2030 a UE[14]. Deși prezentul document se referă în principal la primul scop menționat, importanța pe care etica ar putea să o aibă pentru al doilea scop nu ar trebui să fie subestimată. IA fiabilă poate îmbunătăți prosperitatea individuală și bunăstarea colectivă, generând prosperitate, creând valoare și crescând bogăția. Aceasta poate contribui la realizarea unei societăți echitabile, ajutând la sporirea sănătății și a bunăstării cetățenilor în moduri care să stimuleze egalitatea în ceea ce privește distribuirea oportunităților economice, sociale și politice.

(34) Prin urmare, este absolut necesar să înțelegem cum să sprijinim cel mai bine dezvoltarea, implementarea și utilizarea IA pentru a asigura faptul că toată lumea poate prospera într-o lume bazată pe IA și pentru a construi un viitor mai bun, fiind în același timp competitivi la nivel mondial. La fel ca în cazul oricărei tehnologii puternice, utilizarea sistemelor IA în societatea noastră generează provocări etice, de exemplu legate de impactul acestora asupra oamenilor și societății, asupra capacităților decizionale și siguranței. Dacă vom utiliza tot mai mult asistența sistemelor IA sau deciziile delegate acestora, trebuie să ne asigurăm că sistemele IA sunt echitabile în ceea ce privește impactul pe care îl au asupra vieților oamenilor, că sunt în concordanță cu valori care nu suportă compromisuri și că pot acționa în consecință, precum și că acest lucru poate fi asigurat prin procese adecvate de stabilire a responsabilității.

(35) Europa trebuie să definească ce viziune normativă a unui viitor marcat de omniprezența IA dorește să realizeze și, în consecință, să înțeleagă ce noțiune a IA ar trebui să fie studiată, dezvoltată, implementată și utilizată în Europa pentru a realiza această viziune. Prin prezentul document, intenționăm să contribuim la acest efort, introducând noțiunea de IA fiabilă, care considerăm că este modalitatea corectă de a construi un viitor marcat de IA. Un viitor în care democrația, statul de drept și drepturile fundamentale stau la baza sistemelor IA și în care aceste sisteme îmbunătățesc și apără în mod continuu cultura democratică va face posibil, de asemenea, un mediu în care inovarea și competitivitatea responsabilă pot prospera.

(36) Un cod etic specific domeniului - oricât de coerente, dezvoltate și rafinate ar fi versiunile sale viitoare - nu poate funcționa niciodată ca substitut pentru raționarea etică în sine, care trebuie să rămână întotdeauna sensibilă la detaliile contextuale care nu pot fi cuprinse în orientările generale. Pe lângă dezvoltarea unui set de norme, pentru a asigura IA fiabilă este necesar să construim și să menținem o cultură și o mentalitate etice prin dezbatere publică, educație și învățare practică.

13 https://ec.europa.eu/commission/publications/reflection-paper-towards-sustainable-europe-2030_ro

14 https://sustainabledevelopment.un.org/?menu=1300

1. Drepturile fundamentale ca drepturi morale și legale

(37) Credem într-o abordare a eticii IA care să se bazeze pe drepturile fundamentale consacrate în tratatele UE, în [15]Carta drepturilor fundamentale a Uniunii Europene (Carta UE) și în legislația internațională privind drepturile omului.[16]Respectarea drepturilor fundamentale, în cadrul democrației și al statului de drept, asigură fundamentele cele mai promițătoare pentru a identifica principiile și valorile etice abstracte care pot fi operaționalizate în contextul IA.

(38) Tratatele UE și Carta UE indică o serie de drepturi fundamentale pe care statele membre ale UE și instituțiile UE au obligația legală de a le respecta la punerea în aplicare a dreptului Uniunii. Aceste drepturi sunt descrise în Carta UE prin trimitere la demnitate, libertăți, egalitate și solidaritate, drepturile cetățenilor și justiție. Fundamentul comun care unește aceste drepturi poate fi înțeles ca avându-și originile în respectarea demnității umane - reflectând astfel ceea ce descriem ca „abordare centrată pe om" , în care omul beneficiază de un statut moral unic și inalienabil de supremație în sfera civilă, politică, economică și socială[17].

(39) Deși drepturile stabilite în Carta UE au caracter juridic obligatoriu,[18] este important să se recunoască faptul că drepturile fundamentale nu asigură în fiecare caz o protecție juridică cuprinzătoare. De exemplu, în ceea ce privește Carta UE, este important de subliniat că domeniul său de aplicare este limitat la domeniile dreptului Uniunii. Legislația internațională privind drepturile omului și, în special, Convenția europeană a drepturilor omului au caracter juridic obligatoriu pentru statele membre ale UE, inclusiv în domenii care nu intră în domeniul de aplicare al dreptului Uniunii. În același timp, trebuie subliniat faptul că drepturile fundamentale sunt acordate, de asemenea, persoanelor fizice și (într-o anumită măsură) grupurilor, în temeiul statutului lor moral de ființe umane, în mod independent de forța lor juridică. Prin urmare, fiind înțelese ca drepturi exercitabile legal, drepturile fundamentale se încadrează în prima componentă a IA fiabile (IA legală), care garantează respectarea legii. De asemenea, fiind înțelese ca drepturi ale tuturor, avându-și originile în statutul moral intrinsec al oamenilor, acestea stau la baza celei de a doua componente a IA fiabile (IA etică), ce tratează normele etice care nu au neapărat caracter juridic obligatoriu, dar sunt esențiale pentru asigurarea fiabilității. Deoarece prezentul document nu este menit să ofere îndrumare cu privire la cea dintâi componentă în scopul prezentelor orientări neobligatorii, trimiterile la drepturile fundamentale reflectă cea din urmă componentă.

2. De la drepturile fundamentale la principiile etice

2.1 Drepturile fundamentale ca bază pentru o IA fiabilă

(40) Din setul cuprinzător de drepturi indivizibile stabilite în legislația internațională privind drepturile omului, în Tratatele UE și în Carta UE, familiile de drepturi fundamentale de mai jos sunt deosebit de capabile să acopere sistemele IA. Multe dintre aceste drepturi sunt, în circumstanțele specificate, aplicabile legal în UE, astfel încât respectarea condițiilor acestora este obligatorie din punct de vedere legal. Însă chiar și după ce se asigură respectarea drepturilor fundamentale exercitabile legal, reflecția etică ne poate ajuta să înțelegem modul în care dezvoltarea, implementarea și utilizarea IA pot implica drepturile fundamentale și valorile care stau la baza acestora și pot contribui la asigurarea unei îndrumări mai rafinate atunci când se încearcă să se identifice ce *ar trebui* să facem mai degrabă decât ceea ce *putem* face (în prezent) cu tehnologia.

(41) ***Respectarea demnității umane.*** Demnitatea umană are la bază ideea că fiecare ființă umană deține o „valoare

---

[15] UE se bazează pe un angajament constituțional de a proteja drepturile fundamentale și indivizibile ale oamenilor, de a asigura respectarea statului de drept, de a stimula libertatea democratică și de a promova binele comun. Aceste drepturi sunt reflectate în articolele 2 și 3 din Tratatul privind Uniunea Europeană și în Carta drepturilor fundamentale a UE.

[16] Alte instrumente juridice reflectă și prevăd alte specificații ale acestor angajamente, cum ar fi, de exemplu, Carta socială europeană a Consiliului Europei sau legislația specifică, precum Regulamentul general privind protecția datelor al UE.

[17] Ar trebui remarcat că un angajament față de IA centrată pe om și ancorarea acestuia în drepturile fundamentale necesită fundamente societale și constituționale colective, în care libertatea individuală și respectarea demnității umane sunt atât posibile din punct de vedere practic, cât și semnificative, mai degrabă decât să implice o răspundere individualistă nejustificată a omului.

[18] Conform articolului 51 din Cartă, aceasta se aplică instituțiilor UE și statelor membre ale UE la punerea în aplicare a dreptului Uniunii.

intrinsecă", ce nu ar trebui să fie niciodată diminuată, compromisă sau reprimată de alții - nici de noile tehnologii precum sistemele IA.[19] În contextul IA, respectarea demnității umane implică faptul că toți oamenii sunt tratați cu respectul care li se cuvine în calitate de *subiecți* morali și nu doar de *obiecte* care trebuie selectate, sortate, notate, direcționate prin spirit de turmă, condiționate sau manipulate. Așadar, sistemele IA ar trebui să fie dezvoltate într-o manieră care să respecte, să deservească și să protejeze integritatea fizică și psihică a oamenilor, simțul personal și cultural al identității și satisfacerea nevoilor lor esențiale.[20]

(42) ***Libertatea individului***. Oamenii ar trebui să rămână liberi să adopte decizii de viață pentru ei înșiși. Aceasta implică absența intruziunii suverane, dar necesită, de asemenea, intervenția guvernului și a organizațiilor neguvernamentale pentru a asigura faptul că persoanele fizice sau persoanele expuse riscului de excluziune au acces egal la beneficiile și oportunitățile IA. În contextul IA, libertatea individului necesită atenuarea constrângerii nelegitime (in)directe, a amenințărilor la adresa autonomiei psihice și a sănătății mintale, a supravegherii nejustificate, a inducerii în eroare și a manipulării nedrepte. De fapt, libertatea individului înseamnă angajamentul de a permite persoanelor fizice să exercite un control și mai mare asupra vieților lor, inclusiv (printre alte drepturi), protecția libertății de a desfășura o activitate comercială, libertatea artelor și științelor, libertatea de exprimare, dreptul la viață privată și libertatea de întrunire și de asociere.

(43) ***Respectarea democrației, a justiției și a statului de drept***. Orice putere guvernamentală din democrațiile constituționale trebuie să fie autorizată din punct de vedere legal și limitată prin lege. Sistemele IA ar trebui să servească la menținerea și promovarea sistemelor democratice și la respectarea pluralității valorilor și a alegerilor de viață ale oamenilor. Sistemele IA nu trebuie să submineze procesele democratice, deliberarea umană sau sistemele de vot democratice. De asemenea, sistemele IA trebuie să încorporeze un angajament pentru a asigura faptul că acestea nu funcționează într-un mod care subminează angajamentele fundamentale pe care se întemeiază statul de drept, legile și regulamentele obligatorii și pentru a asigura respectarea garanțiilor procedurale și egalitatea în fața legii.

(44) ***Egalitate, nediscriminare și solidaritate*** - *inclusiv drepturile persoanelor expuse riscului de excluziune*. Trebuie să se asigure respectarea în mod egal a valorii morale și a demnității tuturor oamenilor. Acest concept este mai complex decât nediscriminarea, care tolerează realizarea unor distincții între situații diferite pe baza unor justificări obiective. În contextul IA, egalitatea implică faptul că operațiunile sistemului nu pot genera ieșiri părtinitoare în mod neechitabil (de exemplu, datele utilizate pentru a forma sistemele IA ar trebui să fie cât mai favorabile incluziunii posibil, reprezentând diferite categorii ale populației). Acest lucru necesită, de asemenea, respectarea în mod adecvat a persoanelor și a grupurilor potențial vulnerabile[21], cum ar fi lucrătorii, femeile, persoanele cu handicap, minoritățile etnice, copiii, consumatorii sau alte grupuri expuse riscului de excluziune.

(45) ***Drepturile cetățenilor***. Cetățenii beneficiază de o gamă largă de drepturi, inclusiv dreptul de vot, dreptul la bună administrare sau de acces la documentele publice și dreptul de a adresa petiții administrației. Sistemele IA oferă un potențial substanțial de a îmbunătăți dimensiunea și eficiența guvernului în ceea ce privește furnizarea de bunuri și servicii publice societății. În același timp, drepturile cetățenilor ar putea fi afectate negativ de aplicațiile IA și ar trebui să fie protejate. Utilizarea termenului „drepturile cetățenilor" în prezentul document nu înseamnă negarea sau neglijarea drepturilor resortisanților țărilor terțe și ale persoanelor în situație neregulamentară (sau ilegală) din UE, care au, de asemenea, drepturi conform legislației internaționale și - prin urmare - în domeniul IA.

---

[19] C. McCrudden, *Human Dignity and Judicial Interpretation of Human Rights* (Demnitatea umană și interpretarea judiciară a drepturilor omului), *EJIL*, 19(4), 2008.

[20] Pentru a înțelege „demnitatea umană" în acest sens, a se vedea E. Hilgendorf, *Problem Areas in the Dignity Debate and the Ensemble Theory of Human Dignity* (Domeniile problematice ale dezbaterii privind demnitatea și teoria de ansamblu a demnității umane) din: D. Grimm, A. Kemmerer, C. Möllers (eds.), Human Dignity in Context. Explorations of a Contested Concept (Demnitatea umană în context. Explorările unui concept contestat) 2018, pp. 325 și următoarele.

[21] Pentru o descriere a termenului astfel cum este utilizat în cadrul prezentului document, a se vedea glosarul.

2.2 Principiile etice în contextul sistemelor IA[22]

(46) Numeroase organizații publice, private și civile s-au inspirat din drepturile fundamentale pentru a realiza cadre etice pentru IA.[23] În UE, Grupul european pentru etică în domeniul științei și al noilor tehnologii („EGE") a propus o serie de 9 principii de bază, întemeiate pe valorile fundamentale prevăzute în Tratatele UE și în Carta drepturilor fundamentale a UE.[24] Construim proiectul nostru pe baza acestei activități, recunoscând majoritatea principiilor propuse până în prezent de diverse grupuri, clarificând în același timp scopurile pe care toate principiile urmăresc să le alimenteze și să le sprijine. Aceste principii etice pot reprezenta o sursă de inspirație pentru instrumente de reglementare noi și specifice, pot contribui la interpretarea drepturilor fundamentale, deoarece mediul nostru socio-tehnic evoluează în timp, și pot ghida raționamentul care stă la baza dezvoltării, utilizării și implementării sistemelor IA - adaptându-se în mod dinamic pe măsură ce societatea însăși evoluează.

(47) Sistemele IA ar trebui să îmbunătățească bunăstarea individuală și colectivă. În această secțiune sunt enumerate **patru principii etice** care își au originile în drepturile fundamentale și care trebuie să fie respectate pentru a asigura faptul că sistemele IA sunt dezvoltate, implementate și utilizate în mod fiabil. Acestea sunt specificate ca **imperative etice**, astfel încât practicienii în domeniul IA ar trebui întotdeauna să dorească să le respecte. Fără a impune o ierarhie, în continuare enumerăm principiile într-o manieră care reflectă ordinea în care drepturile fundamentale pe care acestea se bazează sunt menționate în Carta UE.[25]

(48) Acestea sunt principiile:

(i) respectării autonomiei oamenilor;
(ii) prevenirii daunelor;
(iii) echității;
(iv) explicabilității.

(49) Multe dintre acestea sunt reflectate deja într-o mare măsură în cerințele legale existente care trebuie respectate obligatoriu și, prin urmare, intră de asemenea în domeniul de aplicare al „IA legale", care este prima componentă a IA fiabile.[26] Cu toate acestea, astfel cum se prevede mai sus, deși numeroase obligații legale reflectă principiile etice, aderarea la principiile etice nu se limitează la respectarea formală a legislației existente.[27]

- Principiul respectării autonomiei oamenilor

(50) Drepturile fundamentale pe care este întemeiată UE sunt orientate către asigurarea respectării libertății și a autonomiei oamenilor. Oamenii care interacționează cu sistemele IA trebuie să poată să mențină o autodeterminare deplină și eficace asupra lor înșiși și să poată să participe la procesul democratic. Sistemele IA nu ar trebui să subordoneze, să constrângă, să inducă în eroare, să manipuleze, să condiționeze sau să

[22] Aceste principii se aplică și dezvoltării, implementării și utilizării altor tehnologii și, prin urmare, nu sunt specifice sistemelor IA. În cele ce urmează, ne-am propus să prezentăm relevanța lor în mod specific într-un context legat de IA.

[23] De asemenea, recursul la drepturile fundamentale de bază contribuie la limitarea incertitudinii în materie de reglementare, deoarece acesta se poate dezvolta pe baza a decenii de practică a protejării drepturilor fundamentale în UE, oferind astfel claritate, accesibilitate și previzibilitate.

[24] Mai recent, Grupul operativ AI4People a examinat principiile EGE menționate anterior, precum și alte 36 de principii etice stabilite până în prezent și le-a subsumat în cadrul a patru principii generale: L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. J. M. Vayena (2018), „- An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations</Italic>" (AI4People - Un cadru etic pentru o bună societate IA: oportunități, riscuri, principii și recomandări), Minds and Machines 28(4): 689-707.

[25] Respectarea autonomiei oamenilor este puternic asociată cu dreptul la demnitate umană și la libertate (reflectate în articolele 1 și 6 din Cartă). Prevenirea daunelor este strâns legată de protecția integrității fizice sau psihice (reflectată în articolul 3). Echitatea este strâns legată de drepturile la nediscriminare, solidaritate și justiție (reflectate în articolele 21 și următoarele). Explicabilitatea și responsabilitatea sunt strâns legate de drepturile referitoare la justiție (reflectate în articolul 47).

[26] Gândiți-vă, de exemplu, la RGPD sau la regulamentele UE privind protecția consumatorilor.

[27] Pentru o lectură suplimentară cu privire la acest subiect, a se vedea, de exemplu, L. Floridi, *Soft Ethics and the Governance of the Digital* (Etica neobligatorie și guvernanța digitalului), *Philosophy & Technology,* martie 2018, volumul 31, ediția 1, pp 1–8.

direcționeze prin spirit de turmă oamenii în mod nejustificat. În schimb, sistemele IA ar trebui să fie proiectate astfel încât să sporească, să completeze și să permită competențele cognitive, sociale și culturale ale oamenilor. Alocarea funcțiilor între oameni și sistemele IA ar trebui să respecte principii de proiectare centrate pe om și să lase oamenilor o posibilitate semnificativă de alegere. Acest lucru înseamnă asigurarea supravegherii umane[28] și a controlului uman în ceea ce privește proceselor de lucru din sistemele IA. De asemenea, sistemele IA pot modifica în mod fundamental sfera de lucru. Acestea ar trebui să sprijine oamenii în mediul de lucru și să vizeze crearea unei activități semnificative.

- Principiul prevenirii daunelor

(51) Sistemele IA nu ar trebui nici să cauzeze, nici să exacerbeze daunele[29] și nici să afecteze negativ oamenii.[30] Acest lucru implică protecția demnității umane, precum și a integrității psihice și fizice. Sistemele IA și mediile în care acestea funcționează trebuie să fie sigure și securizate. Acestea trebuie să fie solide din punct de vedere tehnic și ar trebui să se asigure faptul că nu pot fi utilizate cu rea intenție. Persoanele vulnerabile ar trebui să primească mai multă atenție și să fie incluse în dezvoltarea și implementarea sistemelor IA. De asemenea, trebuie să se acorde o atenție deosebită situațiilor în care sistemele IA pot să cauzeze sau să exacerbeze efectele negative, din cauza asimetriilor puterii sau ale informațiilor, cum ar fi cele dintre angajatori și angajați, dintre întreprinderi și consumatori sau dintre guverne și cetățeni. De asemenea, prevenirea daunelor implică luarea în considerare a mediului natural și a tuturor ființelor.

- Principiul echității

(52) Dezvoltarea, implementarea și utilizarea sistemelor IA trebuie să fie echitabile. Deși recunoaștem faptul că există numeroase interpretări diferite ale echității, considerăm că echitatea are atât o dimensiune concretă, cât și una procedurală. Dimensiunea concretă implică un angajament de: a asigura distribuirea egală și justă atât a beneficiilor, cât și a costurilor și de a asigura faptul că persoanele fizice și grupurile nu fac obiectul unor judecăți părtinitoare neechitabile, discriminării și stigmatizării. Dacă se pot evita judecățile părtinitoare neechitabile, sistemele IA ar putea chiar să sporească echitatea socială. De asemenea, ar trebui să se promoveze egalitatea de șanse în ceea ce privește accesul la educație, bunuri, servicii și tehnologie. În plus, utilizarea sistemelor IA nu ar trebui să determine niciodată inducerea în eroare sau subminarea utilizatorilor (finali) în ceea ce privește libertatea lor de alegere. În plus, echitatea implică faptul că practicienii în domeniul IA ar trebui să respecte principiul proporționalității dintre mijloace și scopuri și să analizeze cu atenție modul de echilibrare a intereselor și a obiectivelor contrare.[31] Dimensiunea procedurală a echității implică posibilitatea de a contesta și de a ataca în mod eficace deciziile luate de sistemele IA și de oamenii care le utilizează.[32] Pentru a face acest lucru, entitatea responsabilă pentru decizia în cauză trebuie să poată fi identificată, iar procesele decizionale ar trebui să poată fi explicate.

- Principiul explicabilității

(53) Explicabilitatea este esențială pentru stabilirea și menținerea încrederii utilizatorilor în sistemele IA. Acest

28 Conceptul de supraveghere umană este dezvoltat mai mult la punctul 65 de mai jos.

29 Daunele pot fi individuale sau colective și pot include daune intangibile aduse mediului social, cultural și politic.

30 Acest lucru cuprinde, de asemenea, modul de viață al persoanelor fizice și al grupurilor sociale, evitând, de exemplu, daunele culturale.

31 Acest lucru este legat de principiul proporționalității (reflectat în maxima privind „utilizarea unui baros pentru a sparge o nucă"). Măsurile adoptate pentru realizarea unui scop (de exemplu, măsurile de extragere a datelor puse în aplicare pentru a realiza funcția de optimizare a IA) ar trebui să fie limitate la ceea ce este strict necesar. Acest lucru implică, de asemenea, faptul că atunci când mai multe măsuri concurează pentru realizarea unui scop, ar trebui să se acorde prioritate celei care contravine cel mai puțin drepturilor fundamentale și normelor etice (de exemplu, dezvoltatorii din domeniul IA ar trebui să prefere întotdeauna datele din sectorul public față de datele cu caracter personal). De asemenea, se poate face trimitere la proporționalitatea dintre utilizator și organizația de implementare, având în vedere drepturile întreprinderilor (inclusiv proprietatea intelectuală și confidențialitatea), pe de o parte și drepturile utilizatorului, pe de altă parte.

32 Inclusiv făcând uz de dreptul lor de asociere și de a se afilia la sindicate la locul de muncă, astfel cum prevede articolul 12 din Carta drepturilor fundamentale a UE.

lucru înseamnă că procesele trebuie să fie transparente, capacitățile și scopul sistemelor IA trebuie să fie comunicate în mod deschis, iar deciziile - în măsura în care acest lucru este posibil - trebuie să poată fi explicate celor afectați în mod direct sau indirect. Fără aceste informații, o decizie nu poate fi contestată în mod corespunzător. Nu se poate explica întotdeauna de ce un model a generat un anumit rezultat sau o anumită decizie (și ce combinație de factori de intrare a contribuit la acest lucru). Aceste situații sunt denumite algoritmi de tip „black box" (cutie neagră) și necesită o atenție specială. În aceste circumstanțe, pot fi necesare alte măsuri de explicabilitate (de exemplu, trasabilitatea, posibilitatea auditării și comunicarea transparentă cu privire la capacitățile sistemului), cu condiția ca sistemul în ansamblu să respecte drepturile fundamentale. Măsura în care este necesară explicabilitatea depinde foarte mult de contextul și de gravitatea consecințelor, dacă acea ieșire este eronată sau inexactă.[33]

2.3 Tensiunile dintre principii

(54) Între principiile de mai sus pot apărea tensiuni și nu există nicio soluție fixă pentru abordarea acestora. În conformitate cu angajamentul fundamental al UE față de implicarea democratică, respectarea garanțiilor procedurale și participarea politică deschisă, ar trebui să se stabilească metode de deliberare responsabilă pentru a trata aceste tensiuni. De exemplu, în diverse domenii de aplicare, *principiul prevenirii daunelor* și *principiul autonomiei oamenilor* pot fi în conflict. Țineți cont, de exemplu, de utilizarea sistemelor IA pentru „poliția predictivă", care ar putea contribui la reducerea criminalității, dar în moduri care implică activități de supraveghere care afectează libertatea individuală și viața privată. În plus, beneficiile globale ale sistemelor IA ar trebui să depășească în mod substanțial riscurile individuale previzibile. Deși aceste principii oferă cu siguranță îndrumare pentru găsirea unor soluții, ele rămân recomandări etice abstracte. Așadar, nu se poate preconiza că practicienii în domeniul IA vor găsi soluția potrivită pe baza principiilor de mai sus; totuși, aceștia ar trebui să abordeze dilemele și compromisurile etice prin reflecție rațională, bazată pe dovezi, și nu prin intuiție sau alegere aleatorie. Totuși, pot exista situații în care nu pot fi identificate compromisuri acceptabile din punct de vedere etic. Anumite drepturi fundamentale și principii corelate sunt absolute și nu pot face obiectul unui exercițiu de echilibrare (de exemplu, demnitatea umană).

[33] De exemplu, din recomandările inexacte de cumpărături generate de un sistem IA pot rezulta puține preocupări etice, spre deosebire de sistemele IA care evaluează dacă o persoană condamnată pentru o infracțiune ar trebui să fie eliberată condiționat.

Orientări esențiale rezultate din capitolul I:

- ✓ Dezvoltarea, implementarea și utilizarea sistemelor IA astfel încât să se respecte principiile etice ale: *respectării autonomiei oamenilor, prevenirii daunelor, echității și explicabilității.* Recunoașterea și abordarea tensiunilor potențiale dintre aceste principii.
- ✓ Acordarea unei atenții deosebite situațiilor care implică mai multe grupuri vulnerabile, cum ar fi copiii, persoanele cu handicap și alte grupuri care au fost defavorizate la nivel istoric, care sunt expuse riscului de excluziune și/sau situațiilor caracterizate de asimetrii ale puterii sau ale informațiilor, cum ar fi cele dintre angajatori și lucrători sau dintre întreprinderi și consumatori[34].
- ✓ Recunoașterea și luarea în considerare a faptului că, deși au potențialul de a aduce numeroase beneficii substanțiale persoanelor fizice și societății, unele aplicații IA pot avea totodată un impact negativ, inclusiv un impact care poate fi dificil de anticipat, de identificat sau de măsurat (de exemplu, asupra democrației, statului de drept și justiției distributive sau chiar asupra minții umane). Adoptarea unor măsuri adecvate pentru abordarea acestor riscuri dacă este cazul și în mod proporțional cu magnitudinea riscului.

II. Capitolul II: Asigurarea IA fiabile

(55) Acest capitol oferă orientări privind punerea în aplicare și realizarea IA fiabile, prin intermediul unei liste de șapte cerințe care ar trebui să fie îndeplinite și care se bazează pe principiile prezentate în capitolul I. În plus, sunt introduse metodele tehnice și metodele fără caracter tehnic disponibile în prezent pentru punerea în aplicare a acestor cerințe pe toată durata ciclului de viață al sistemului IA.

1. Cerințele pentru o IA fiabilă

(56) Principiile prezentate în capitolul I trebuie să fie traduse în cerințe concrete pentru a realiza o IA fiabilă. Aceste cerințe se aplică diferitelor părți interesate care participă la ciclul de viață al sistemelor IA: dezvoltatorii, organizațiile de implementare și utilizatorii finali, precum și societatea în general. Dezvoltatorii sunt cei care cercetează, proiectează și/sau dezvoltă sistemele IA. Organizații de implementare înseamnă organizațiile publice sau private care utilizează sistemele IA în cadrul proceselor lor de afaceri și pentru a oferi produse și servicii altora. Utilizatorii finali sunt cei care interacționează cu sistemul IA, în mod direct sau indirect. În final, societatea în general cuprinde toate celelalte părți care sunt afectate în mod direct sau indirect de sistemele IA.

(57) Diferitele clase de părți interesate trebuie să joace diferite roluri pentru a asigura îndeplinirea cerințelor:

a. dezvoltatorii ar trebui să implementeze și să aplice cerințele în procesele de proiectare și dezvoltare;
b. organizațiile de implementare ar trebui să asigure faptul că sistemele pe care le utilizează și produsele și serviciile pe care le oferă îndeplinesc cerințele;
c. utilizatorii finali și societatea în general ar trebui să fie informați cu privire la aceste cerințe și să poată solicita ca acestea să fie menținute.

(58) Lista de cerințe de mai jos este neexhaustivă[35]. Aceasta include aspecte sistemice, individuale și societale:

1 Factorul uman și supravegherea umană

Inclusiv drepturile fundamentale, factorul uman și supravegherea umană

2 Soliditatea tehnică și siguranța

Inclusiv reziliența la atacuri și securitatea, soluțiile de siguranță și siguranța generală, acuratețea,

[34] A se vedea articolele 24-27 din Carta UE, care consacră drepturile copilului și ale persoanelor în vârstă, integrarea persoanelor cu handicap și drepturile lucrătorilor. A se vedea, de asemenea, articolul 38 care consacră protecția consumatorilor.

[35] Fără a impune o ierarhie, în continuare enumerăm principiile într-o manieră care reflectă ordinea în care principiile și drepturile la care se referă sunt menționate în Carta UE.

fiabilitatea și reproductibilitatea

3 **Viața privată și guvernanța datelor**

Inclusiv respectarea vieții private, calitatea și integritatea datelor și accesul la date

4 **Transparență**

Inclusiv trasabilitatea, explicabilitatea și comunicarea

5 **Diversitate, nediscriminare și echitate**

Inclusiv evitarea judecăților părtinitoare neechitabile, accesibilitatea și proiectarea universală, precum și participarea părților interesate

6 **Bunăstarea socială și ecologică**

Inclusiv durabilitatea și caracterul ecologic, impactul social, societatea și democrația

7 **Responsabilitate**

Inclusiv posibilitatea auditării, reducerea la minimum și raportarea impactului negativ, compromisurile și măsurile de reparare.

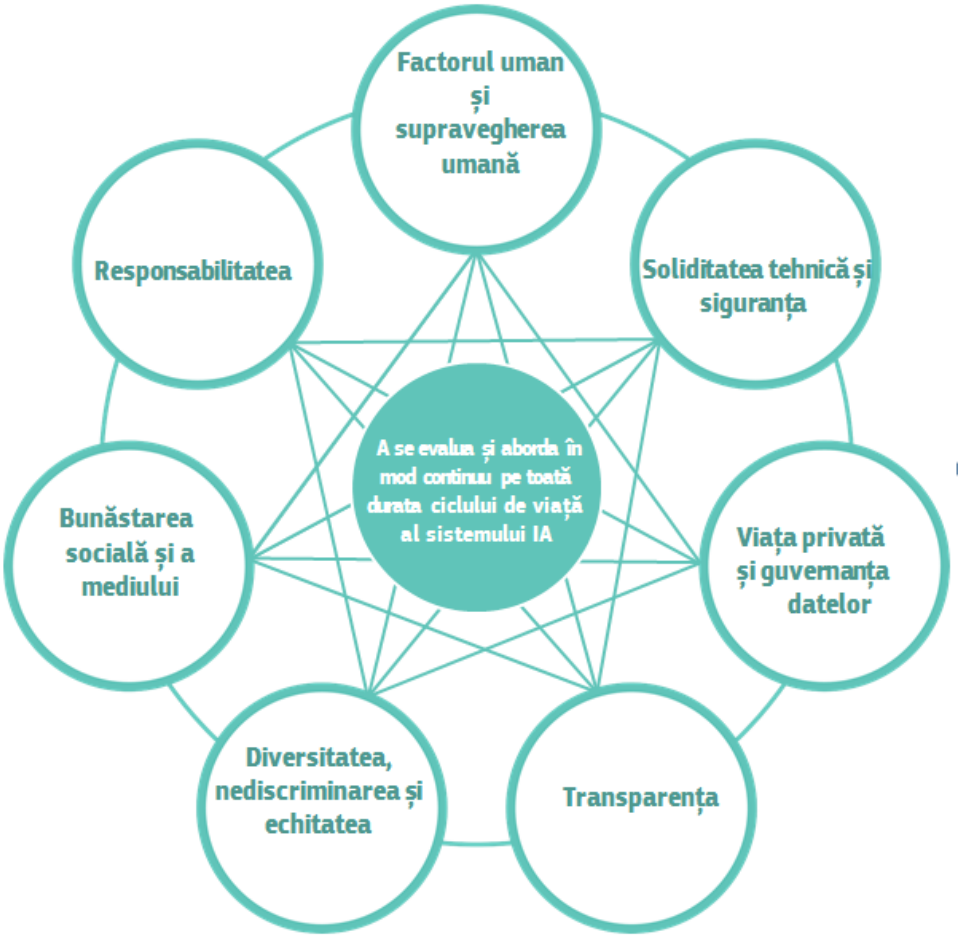

Figura 2: Interdependența celor șapte cerințe: toate sunt la fel de importante, se sprijină reciproc și ar trebui să fie puse în aplicare și evaluate pe toată durata ciclului de viață al sistemului IA

(59) Deși cerințele sunt la fel de importante, contextul și potențialele tensiuni dintre acestea vor trebui să fie luate în considerare la aplicarea acestora în domenii și industrii diferite. Punerea în aplicare a acestor cerințe ar trebui să aibă loc pe toată durata ciclului de viață al sistemului IA și depinde de aplicația specifică. Deși majoritatea cerințelor se aplică tuturor sistemelor IA, se acordă atenție specială celor care afectează persoanele fizice în mod direct sau indirect. Prin urmare, pentru unele aplicații (de exemplu, în mediile industriale), acestea pot avea mai puțină relevanță.

(60) Cerințele de mai sus includ elemente care, în unele cazuri, sunt reflectate deja în legile existente. Reiterăm faptul că - în conformitate cu prima componentă a IA fiabile - este responsabilitatea dezvoltatorilor și a organizațiilor de implementare a sistemelor IA să asigure faptul că acestea îndeplinesc obligațiile lor legale, atât în ceea ce privește normele aplicabile la nivel orizontal, cât și regulamentele specifice domeniului.

(61) La punctele următoare, fiecare cerință va fi explicată în mod detaliat.

1. <u>Factorul uman și supravegherea umană</u>

(62) Sistemele IA ar trebui să sprijine autonomia oamenilor și luarea deciziilor, conform principiului *respectării autonomiei oamenilor*. În acest sens, este nevoie atât ca sistemele IA să acționeze ca motoare pentru o societate democratică, înfloritoare și echitabilă, sprijinind deciziile utilizatorului, cât și să promoveze drepturile fundamentale și să permită supravegherea de către om.

(63) ***Drepturile fundamentale***. La fel ca multe tehnologii, sistemele IA pot să capaciteze și, în egală măsură, să obstrucționeze drepturile fundamentale. De exemplu, acestea pot fi benefice pentru oameni, ajutându-i să își identifice datele cu caracter personal sau sporind accesibilitatea educației, sprijinind, așadar, dreptul lor la educație. Totuși, având în vedere domeniul de aplicare și capacitatea sistemelor IA, acestea pot, de asemenea, să afecteze negativ drepturile fundamentale. În situațiile în care există astfel de riscuri, ar trebui să se efectueze o evaluare a impactului asupra drepturilor fundamentale. Aceasta ar trebui să se efectueze înainte de dezvoltarea sistemelor IA și să includă o evaluare prin care să se stabilească dacă riscurile pot fi reduse sau justificate ca fiind necesare într-o societate democratică în vederea respectării drepturilor și libertăților altor persoane. În plus, ar trebui să se instituie mecanisme pentru a primi feedback extern cu privire la sistemele IA care ar putea încălca drepturile fundamentale.

(64) ***Factorul uman.*** Utilizatorii ar trebui să poată lua decizii autonome informate cu privire la sistemele IA. Acestora ar trebui să li se pună la dispoziție cunoștințele și instrumentele necesare pentru a înțelege sistemele IA și a interacționa cu ele într-un mod satisfăcător și, dacă este posibil, să li se permită să autoevalueze în mod rezonabil sau să conteste sistemul. Sistemele IA ar trebui să ajute persoanele să ia decizii mai bune și mai bine informate, în conformitate cu propriile obiective. Uneori, sistemele IA pot fi implementate astfel încât să modeleze și să influențele comportamentul uman prin mecanisme care pot fi dificil de detectat, deoarece ele pot valorifica procesele de la nivelul subconștientului, inclusiv prin diverse forme de manipulare neechitabilă, inducere în eroare, direcționare prin spirit de turmă și condiționare, toate acestea putând amenința autonomia individuală. Principiul general al autonomiei utilizatorului ar trebui se afle în centrul funcționării sistemului. În acest scop este esențial dreptul de a nu face obiectul unei decizii bazate exclusiv pe prelucrarea automată atunci când aceasta produce efecte juridice asupra utilizatorilor sau îi afectează în mod similar într-o măsură semnificativă[36].

(65) ***Supravegherea umană***. Supravegherea umană contribuie la asigurarea faptului că sistemele IA nu subminează autonomia umană și nu provoacă alte efecte adverse. Supravegherea se poate realiza prin mecanisme de guvernanță, precum asigurarea unei abordări care implică factorul uman, de tipul human-in-the-loop (omul care este implicat în operațiune - HITL), human-on-the-loop (omul care monitorizează operațiunea - HOTL) sau human-in-command (omul care conduce operațiunea - HIC). Abordarea de tipul HITL se referă la capacitatea de intervenție umană în fiecare ciclu decizional al sistemului, care, în numeroase situații, nu este nici posibilă, nici de dorit. Abordarea de tipul HOTL referă la capacitatea de intervenție umană în ciclul de proiectare a sistemului și la monitorizarea funcționării sistemului. Abordarea de tipul HIC se referă la capacitatea de a supraveghea activitatea generală a sistemului IA (inclusiv a impactului său economic, societal, juridic și etic mai amplu) și la capacitatea de a decide când și cum se folosește sistemul în orice situație specifică. Aici poate intra decizia de a nu folosi un sistem IA într-o situație specifică, de a stabili niveluri de discreție umană pe perioada de utilizare a sistemului sau de a asigura capacitatea de a lua o decizie care să o înlocuiască pe cea luată de

[36] Se poate face trimitere la articolul 22 din RGPD, unde acest drept este deja consacrat.

sistem. În plus, trebuie să se asigure faptul că autoritățile publice de aplicare a legii au capacitatea de a-și exercita atribuțiile de supraveghere în conformitate cu mandatele lor. Mecanismele de supraveghere pot fi necesare într-o măsură mai mică sau mai mare pentru a sprijini alte măsuri de siguranță și control, în funcție de domeniul de aplicare și de riscul potențial al sistemului IA. Dacă toate celelalte condiții rămân neschimbate, cu cât o ființă umană poate exercita mai puțină supraveghere asupra unui sistem IA, cu atât este nevoie de teste mai extinse și de o guvernanță mai strictă.

2. <u>Soliditatea tehnică și siguranța</u>

(66) O componentă esențială pentru realizarea IA fiabile este soliditatea tehnică, strâns legată de *principiul prevenirii daunelor*. Soliditatea tehnică prevede ca sistemele IA să fie dezvoltate cu o abordare preventivă a riscurilor, astfel încât acestea să se comporte în mod fiabil conform scopului, reducând totodată la minimum daunele neintenționate și neașteptate și prevenind daunele inacceptabile. Acest lucru ar trebui să se aplice, de asemenea, potențialelor modificări ale mediului lor de funcționare sau prezenței altor agenți (umani și artificiali) care pot interacționa cu sistemul într-o manieră contradictorie. În plus, ar trebui să se asigure integritatea fizică și psihică a oamenilor.

(67) ***Reziliența la atacuri și securitatea.*** Sistemele IA, ca toate sistemele software, ar trebui să fie protejate împotriva vulnerabilităților care pot permite exploatarea de către adversari, de exemplu pirateria informatică. Atacurile pot viza datele (*data poisoning*), modelul (scurgere de modele) sau infrastructura subiacentă, atât software-ul, cât și hardware-ul. Dacă un sistem IA este atacat, de exemplu în cadrul unor atacuri adversare, datele, precum și comportamentul sistemului pot fi modificate, ceea ce determină sistemul să adopte decizii diferite sau să se închidă cu totul. De asemenea, sistemele și datele pot fi corupte prin rea intenție sau prin expunerea la situații neașteptate. Procesele de securitate insuficiente pot, de asemenea, să determine decizii eronate sau chiar prejudicii fizice. Pentru ca sistemele IA să fie considerate sigure,[37] ar trebui să se ia în considerare posibilele aplicații neintenționate ale IA (de exemplu, aplicațiile cu dublă utilizare) și abuzarea potențială a unui sistem IA de către actori rău intenționați; ar trebui, astfel, să se adopte măsuri pentru a preveni și a atenua aceste lucruri.[38]

(68) ***Soluții de siguranță și siguranța generală.*** Sistemele IA ar trebui să aibă dispozitive de securitate care să permită soluții de siguranță în cazul în care există probleme. Acest lucru poate însemna că sistemele IA trec de la o procedură statistică la una bazată pe reguli sau că solicită un operator uman înainte de a-și continua acțiunea.[39] Trebuie să se asigure faptul că sistemul face ceea ce ar trebui să facă fără să producă daune ființelor sau mediului. Acest lucru include reducerea la minimum a consecințelor neintenționate și a erorilor. În plus, ar trebui instituite procese de clarificare și de evaluare a riscurilor potențiale asociate utilizării sistemelor IA, în diferite domenii de aplicare. Nivelul de măsuri de siguranță necesar depinde de magnitudinea riscului pe care îl prezintă un sistem IA, care, la rândul său, depinde de capacitățile sistemului. În cazul în care se poate preconiza că procesul de dezvoltare sau sistemul în sine va prezenta riscuri deosebit de mari, este esențial să se dezvolte și să se testeze în mod proactiv măsuri de siguranță.

(69) ***Acuratețea.*** Acuratețea se referă la capacitatea unui sistem IA de a face aprecieri corecte, de exemplu, de a clasifica informațiile în mod corect în categorii adecvate sau capacitatea acestuia de a face previziuni, recomandări corecte sau de a lua decizii pe baza datelor sau a modelelor. Un proces explicit și bine format de dezvoltare și de evaluare poate să sprijine, să atenueze și să corecteze riscurile neintenționate care rezultă din

[37] A se vedea, de exemplu, considerentele de la punctul 2.7 din Planul coordonat privind inteligența artificială al Uniunii Europene.

[38] Pentru securitatea sistemelor IA, poate fi extrem de necesar să se poată realiza un cerc virtuos în cercetare și dezvoltare, în ceea ce privește înțelegerea atacurilor, dezvoltarea unor protecții adecvate și îmbunătățirea metodologiilor de evaluare. Pentru a realiza acest lucru, ar trebui să se promoveze o convergență între comunitatea IA și comunitatea de securitate. În plus, este responsabilitatea tuturor actorilor implicați să creeze norme comune de siguranță și securitate transfrontalieră și să stabilească un mediu de încredere reciprocă, promovând colaborarea internațională. Pentru măsurile posibile, a se vedea *Malicious Use of AI* (Utilizarea rău intenționată a IA) (Avin S., Brundage M., et. al., 2018).

[39] De asemenea, ar trebui să se ia în considerare scenariile în care intervenția umană nu ar fi posibilă imediat.

previziuni incorecte. Atunci când previziunile inexacte ocazionale nu pot fi evitate, este important ca sistemul să poată indica gradul de probabilitate al acestor erori. Un nivel înalt de acuratețe este esențial în special în situațiile în care sistemul IA afectează în mod direct viețile oamenilor.

(70) ***Fiabilitatea și reproductibilitatea.*** Este esențial ca rezultatele sistemelor IA să fie reproductibile, precum și fiabile. Un sistem IA fiabil este un sistem care funcționează în mod corespunzător cu o gamă de intrări și într-o serie de situații. Acest lucru este necesar pentru a verifica sistemul IA și pentru a preveni daunele neintenționate. Reproductibilitatea descrie dacă un experiment IA prezintă același comportament atunci când este repetat în aceleași condiții. Acest lucru permite cercetătorilor științifici și factorilor de decizie să descrie cu exactitate ceea ce fac sistemele IA. Fișierele de replicare[40] pot facilita procesul de testare și de reproducere a comportamentelor.

3. Viața privată și guvernanța datelor

(71) Viața privată, un drept fundamental afectat de sistemele IA, este strâns legată de *principiul prevenirii daunelor*. Prevenirea daunelor asupra vieții private necesită, de asemenea, o guvernanță adecvată a datelor, care să acopere calitatea și integritatea datelor utilizate, relevanța acestora ținând seama de domeniul în care vor fi implementate sistemele IA, protocoalele de acces la acestea și capacitatea de a prelucra datele astfel încât să se protejeze viața privată.

(72) *Viața privată și protecția datelor* Sistemele IA trebuie să garanteze protecția vieții private și a datelor pe toată durata ciclului de viață al sistemului.[41] Acest lucru include informațiile furnizate inițial de utilizator, precum și informațiile generate cu privire la utilizator pe parcursul interacțiunii acestuia cu sistemul (de exemplu, ieșirile pe care sistemul IA le-a generat pentru utilizatori specifici sau modul în care utilizatorii au răspuns anumitor recomandări). Înregistrările digitale ale comportamentului uman pot permite sistemelor IIA să deducă nu numai preferințele persoanelor fizice, ci și orientarea sexuală, vârsta, genul, opiniile religioase sau politice ale acestora. Pentru ca persoanele fizice să poată avea încredere în procesul de colectare a datelor, trebuie să se asigure faptul că datele colectate cu privire la acestea nu vor fi utilizate pentru a le discrimina în mod ilegal sau inechitabil.

(73) ***Calitatea și integritatea datelor***. Calitatea seturilor de date utilizate este fundamentală pentru performanța sistemelor IA. Atunci când sunt colectate datele, acestea pot conține judecăți părtinitoare determinate social, inexactități, erori și greșeli. Aceste situații trebuie abordate înainte de formarea pe baza unui set de date furnizat. În plus, trebuie asigurată integritatea datelor. Alimentarea unui sistem IA cu date periculoase poate schimba comportamentul acestuia, în special în cazul sistemelor care au capacitate de auto-învățare. Procesele și seturile de date utilizate trebuie să fie testate și documentate în fiecare etapă, cum ar fi în etapele de planificare, formare, testare și implementare. Această dispoziție ar trebui să se aplice și în cazul sistemelor IA care nu au fost dezvoltate intern, ci dobândite din altă sursă.

(74) ***Accesul la date***. În orice organizație care manipulează date ale persoanelor fizice (indiferent dacă o persoană este sau nu utilizator al sistemului), ar trebui să existe protocoale de date care să guverneze accesul la date. Aceste protocoale ar trebui să definească cine poate accesa datele și în ce circumstanțe. Doar personalului calificat în mod corespunzător, care are competența și necesitatea de a accesa datele persoanei fizice, ar trebui să i se permită acest lucru.

4. Transparență

[40] Acest lucru se referă la fișierele care pot replica fiecare etapă a procesului de dezvoltare a sistemului IA, de la cercetare și colectarea datelor inițiale până la rezultate.

[41] Se poate face trimitere la legile existente privind viața privată, cum ar fi RGPD sau viitorul Regulament privind viața privată și comunicațiile electronice.

(75) Această cerință este strâns legată de *principiul explicabilității* și cuprinde transparența elementelor relevante pentru un sistem IA: datele, sistemul și modelele de afaceri.

(76) ***Trasabilitate***. Seturile de date și procesele care determină decizia sistemului IA, inclusiv cele de colectare și de etichetare a datelor, precum și algoritmii utilizați ar trebui să fie documentați la cel mai bun standard posibil pentru a permite trasabilitatea și o creștere a transparenței. Această prevedere se aplică, de asemenea, pentru deciziile adoptate de sistemul IA. Acest lucru permite identificarea motivelor pentru care o decizie din domeniul IA a fost eronată, ceea ce, la rândul său, ar putea contribui la prevenirea greșelilor viitoare. Prin urmare, trasabilitatea facilitează posibilitatea de auditare și explicabilitatea.

(77) ***Explicabilitatea.*** Explicabilitatea se referă la capacitatea de a explica atât procesele tehnice ale unui sistem IA, cât și deciziile umane asociate (de exemplu, domeniile de aplicare ale unui sistem IA). Explicabilitatea tehnică prevede ca deciziile adoptate de un sistem IA să poată fi înțelese și identificate de oameni. În plus, ar putea fi necesare compromisuri între consolidarea explicabilității unui sistem (care îi poate reduce acuratețea) sau sporirea acurateței acestuia (în detrimentul explicabilității). Ori de câte ori un sistem IA are un impact semnificativ asupra vieților oamenilor, ar trebui să fie posibil să se solicite o explicație adecvată pentru procesul decizional al sistemului IA. Această explicație ar trebui să fie oportună și adaptată expertizei părții interesate în cauză (de exemplu, un nespecialist, o autoritate de reglementare sau un cercetător). În plus, ar trebui furnizate explicații privind măsura în care un sistem IA influențează și modelează procesul decizional organizațional, opțiunile în ceea ce privește proiectarea sistemului, precum și raționamentul care stă la baza implementării acestuia (asigurându-se, astfel, transparența modelului de afaceri).

(78) ***Comunicarea.*** Sistemele IA nu ar trebui să se prezinte drept oameni în fața utilizatorilor; oamenii au dreptul de a fi informați cu privire la faptul că interacționează cu un sistem IA. Acest lucru presupune faptul că sistemele IA trebuie să poată fi identificate ca atare. În plus, ar trebui să se ofere posibilitatea de a decide împotriva acestei interacțiuni și în favoarea interacțiunii umane, atunci când este necesar, pentru a asigura respectarea drepturilor fundamentale. În plus, capacitățile și limitările sistemului IA ar trebui să fie comunicate practicienilor în domeniul IA sau utilizatorilor finali într-o manieră adecvată utilizării în cauză. Acest lucru ar putea cuprinde comunicarea nivelului de acuratețe al sistemului IA, precum și limitările acestuia.

5. Diversitate, nediscriminare și echitate

(79) Pentru o dezvolta o IA fiabilă, trebuie să permitem incluziunea și diversitatea pe toată durata ciclului de viață al sistemului IA. Pe lângă luarea în considerare și implicarea tuturor părților interesate afectate pe toată durata procesului, acest lucru presupune, de asemenea, asigurarea accesului egal prin procese de proiectare favorabile incluziunii, precum și prin egalitate de tratament. Această cerință este strâns legată de *principiul echității*.

(80) ***Evitarea judecăților părtinitoare neechitabile***. Seturile de date folosite de sistemele IA (atât pentru formare, cât și pentru funcționare) pot include judecăți părtinitoare istorice neintenționate, date incomplete și modele de guvernanță deficitară. Perpetuarea acestor judecăți părtinitoare ar putea conduce la prejudicii și discriminare neintenționată (in)directă[42] față de anumite grupuri sau persoane, existând posibilitatea de a se exacerba prejudiciile sau marginalizarea. De asemenea, pot rezulta daune și din exploatarea intenționată a judecăților părtinitoare (ale consumatorilor) sau prin realizarea unor acte de concurență neloială, cum ar fi omogenizarea prețurilor prin coluziune sau o piață netransparentă.[43] Judecățile părtinitoare identificabile și

[42] Pentru o definiție a discriminării directe și indirecte, a se vedea, de exemplu, articolul 2 din Directiva 2000/78/CE a Consiliului din 27 noiembrie 2000 de creare a unui cadru general în favoarea egalității de tratament în ceea ce privește încadrarea în muncă și ocuparea forței de muncă. A se vedea, de asemenea, articolul 21 din Carta drepturilor fundamentale a UE.

[43] A se vedea documentul Agenției pentru Drepturi Fundamentale a Uniunii Europene: *„BigData: Discrimination in data-supported decision making* (2018)" (Big Data: discriminarea în luarea deciziilor pe bază de date) http://fra.europa.eu/en/publication/2018/big-data-discrimination

discriminatorii ar trebui să fie eliminate în etapa de colectare, dacă este posibil. De asemenea, modul în care sunt dezvoltate sistemele IA (de exemplu, programarea algoritmilor) poate suferi de pe urma judecăților părtinitoare neechitabile. Acest lucru ar putea fi contracarat prin instituirea unor procese de supraveghere pentru a analiza și a aborda scopul sistemului, constrângerile, cerințele și deciziile într-un mod clar și transparent. În plus, angajarea unor persoane din medii, culturi și discipline diverse poate asigura diversitatea de opinii și ar trebui să fie încurajată.

(81) ***Accesibilitate și proiectarea universală***. În special în domeniile business-to-consumer, sistemele ar trebui să fie centrate pe utilizator și proiectate astfel încât să permită tuturor persoanelor să utilizeze produsele sau serviciile IA, indiferent de vârstă, gen, abilități sau caracteristici. Accesibilitatea la această tehnologie pentru persoanele cu handicap, care sunt prezente în toate grupurile sociale, este deosebit de importantă. Sistemele IA nu ar trebui să aibă o abordare nediferențiată și ar trebui să ia în considerare principiile proiectării universale[44], care abordează cea mai largă gamă posibilă de utilizatori, respectând standardele de accesibilitate relevante.[45] Acest lucru va permite accesul echitabil și participarea activă a tuturor persoanelor la activitățile umane existente și emergente mediate de computer și cu privire la tehnologiile de asistență.[46]

(82) ***Participarea părților interesate***. Pentru a dezvolta sisteme IA fiabile, se recomandă consultarea părților interesate care pot fi afectate în mod direct sau indirect de sistem pe toată durata ciclului său de viață. Este benefic să se solicite feedback periodic, chiar și după implementare, și să se instituie mecanisme pe termen lung pentru participarea părților interesate, de exemplu asigurând informarea, consultarea și participarea lucrătorilor pe parcursul întregului proces de punere în aplicare a sistemelor IA în cadrul organizațiilor.

6. Bunăstare socială și ecologică

(83) În conformitate cu *principiile echității* și *prevenirii daunelor*, societatea în general, alte ființe sensibile și mediul ar trebui să fie luate în considerare ca părți interesate pe toată durata ciclului de viață al IA. Durabilitatea și responsabilitatea ecologică a sistemelor IA ar trebui să fie încurajate, iar cercetarea ar trebui să fie promovată în cadrul soluțiilor IA care abordează domenii de interes global, cum ar fi, de exemplu, obiectivele de dezvoltare durabilă. În mod ideal, IA ar trebui să fie utilizată în beneficiul tuturor ființelor umane, inclusiv al generațiilor viitoare.

(84) ***IA durabilă și ecologică***. Sistemele IA promit să contribuie la abordarea unora dintre cele mai presante preocupări sociale și, totuși, trebuie să se asigure faptul că acest lucru are loc în modul cel mai ecologic posibil. În această privință, procesul de dezvoltare, implementare și utilizare a sistemului, precum și întregul său lanț de aprovizionare ar trebui să fie evaluate, de exemplu, printr-o examinare critică a utilizării resurselor și a consumului de energie pe durata formării, optând pentru soluții mai puțin dăunătoare. Ar trebui să fie încurajate măsurile de asigurare a caracterului ecologic al întregului lanț de aprovizionare al sistemului IA.

(85) ***Impactul social***. Expunerea omniprezentă la sistemele IA sociale[47] în toate domeniile vieților noastre (fie în educație, muncă, asistență sau divertisment) poate să ne modifice concepția despre factorul social sau să ne afecteze relațiile sociale și atașamentul. Deși sistemele IA pot fi utilizate pentru dezvoltarea competențelor sociale,[48] acestea pot contribui și la deteriorarea lor. Acest lucru poate afecta, de asemenea, bunăstarea fizică

[44] Articolul 42 din Directiva privind achizițiile publice prevede ca specificațiile tehnice să ia în considerare accesibilitatea și proiectarea pentru toate categoriile de utilizatori.

[45] De exemplu, EN 301 549.

[46] Această cerință are legătură cu Convenția Națiunilor Unite privind drepturile persoanelor cu handicap.

[47] Aceasta vizează sistemele IA care comunică și interacționează cu oamenii, simulând socialitatea în interacțiunea dintre om și robot (IA integrată), sau ca avataruri în realitatea virtuală. Astfel, acele sisteme au potențialul de a schimba practicile noastre socio-culturale și structura vieții noastre sociale.

[48] A se vedea, de exemplu, proiectul finanțate de UE pentru dezvoltarea de software pe bază de IA, care permite roboților să interacționeze în mod mai eficace cu copiii care suferă de autism în cadrul ședințelor de terapie conduse de oameni, ajutând la îmbunătățirea abilităților lor sociale și de comunicare:

și psihică a oamenilor. Prin urmare, efectele acestor sisteme trebuie să fie monitorizate și luate în considerare cu atenție.

(86) ***Societate și democrație.*** Pe lângă evaluarea impactului pe care dezvoltarea, implementarea și utilizarea unui sistem IA îl au asupra persoanelor fizice, acest impact ar trebui să fie evaluat și din perspectivă socială, luând în considerare efectul său asupra instituțiilor, democrației și societății în general. Utilizarea sistemelor IA ar trebui luată în considerare în mod corespunzător, în special în situații ce au legătură cu procesul democratic, nu numai cu luarea deciziilor politice, ci și în contexte electorale.

7. Responsabilitate

(87) Cerința privind responsabilitatea completează cerințele de mai sus, fiind strâns legată de *principiul echității*. Aceasta prevede că este necesară instituirea unor mecanisme pentru asigurarea responsabilității și a asumării răspunderii pentru sistemele IA și rezultatele acestora, atât înainte de punerea lor în aplicare, cât și ulterior.

(88) ***Posibilitatea auditării.*** Posibilitatea auditării implică permiterea evaluării algoritmilor, a datelor și a proceselor de proiectare. Acest lucru nu presupune neapărat că informațiile cu privire la modelele de afaceri și proprietatea intelectuală legate de sistemul IA trebuie să fie întotdeauna disponibile în mod liber. Evaluarea de către auditori interni și externi și disponibilitatea acestor rapoarte de evaluare pot contribui la fiabilitatea tehnologiei. În aplicațiile care afectează drepturile fundamentale, inclusiv în aplicațiile critice din punctul de vedere al siguranței, sistemele IA ar trebui să poată fi auditate în mod independent.

(89) ***Reducerea la minimum și raportarea impactului negativ***. Trebuie să se asigure atât capacitatea de a raporta cu privire la acțiunile sau deciziile care contribuie la un anumit rezultat al sistemului, cât și cea de a răspunde la consecințele unui astfel de rezultat. Identificarea, evaluarea, raportarea și reducerea la minimum a impactului negativ potențial al sistemelor IA sunt esențiale în mod special pentru cei afectați în mod (in)direct. Trebui să fie disponibilă o protecție corespunzătoare pentru avertizori, ONG-uri, sindicate sau alte entități atunci când raportează preocupări legitime cu privire la un sistem bazat pe IA. Utilizarea evaluărilor impactului (de exemplu, Red Teaming sau forme de evaluare algoritmică a impactului), atât înainte de dezvoltarea, implementarea și utilizarea sistemelor IA, cât și ulterior, poate fi utilă pentru reducerea la minimum a impactului negativ. Aceste evaluări trebuie să fie proporționale cu riscul pe care îl prezintă sistemele IA.

(90) ***Compromisuri.*** La punerea în aplicare a cerințelor de mai sus, pot apărea tensiuni între acestea, care pot conduce la compromisuri inevitabile. Aceste compromisuri ar trebui să fie abordate într-un mod rațional și metodologic ținând cont de stadiul actual al tehnologiei. Acest lucru presupune faptul că ar trebui să fie identificate interesele și valorile relevante pe care le implică sistemul IA și că, dacă apar conflicte, compromisurile dintre ele ar trebui să fie recunoscute în mod explicit și evaluate în ceea ce privește riscul pe care acestea îl prezintă pentru principiile etice, inclusiv pentru drepturile fundamentale. În situațiile în care nu pot fi identificate compromisuri acceptabile din punct de vedere etic, dezvoltarea, implementarea și utilizarea sistemului IA nu ar trebui să se desfășoare în această formă. Orice decizie cu privire la compromisul care trebuie făcut ar trebui să fie motivată și documentată în mod adecvat. Factorul de decizie trebuie să fie răspunzător pentru modul în care se face compromisul adecvat și ar trebui să revizuiască în mod continuu caracterul oportun al deciziei rezultate, pentru a asigura faptul că se pot efectua modificările necesare ale sistemului, dacă este necesar. [49]

---

http://ec.europa.eu/research/infocentre/article_en.cfm?id=/research/headlines/news/article_19_03_12_en.html?infocentre&item=Infocentre&artid=49968

[49] La realizarea acestui lucru pot contribui diferite modele de guvernanță. De exemplu, prezența unui expert sau a unui consiliu etic (și specific sectorului) intern și/sau extern ar putea fi utilă pentru a evidenția domeniile de conflict potențial și pentru a sugera modalități în care conflictul respectiv ar putea fi soluționat cel mai bine. De asemenea, este utilă o consultare semnificativă și o discuție cu părțile interesate, inclusiv cu cele expuse riscului de a fi afectate negativ de un sistem IA. Universitățile europene ar trebui să își asume un rol principal în formarea experților în etică necesari.

(91) ***Măsuri de reparare***. Atunci când se produce un impact negativ injust, ar trebui prevăzute mecanisme accesibile, care să asigure măsuri de reparare adecvate[50]. Pentru a asigura încrederea, este esențial să se cunoască faptul că sunt posibile măsuri de reparare atunci când apar probleme. Ar trebui să se acorde o atenție deosebită persoanelor sau grupurilor vulnerabile.

2. Metodele tehnice și metodele fără caracter tehnic pentru realizarea unei IA fiabile

(92) Pentru a pune în aplicare cerințele de mai sus, pot fi utilizate atât metode tehnice, cât și metode fără caracter tehnic. Acestea cuprind toate etapele ciclului de viață al unui sistem IA. O evaluare a metodelor utilizate pentru punerea în aplicare a cerințelor, precum și raportarea și justificarea[51] modificărilor aduse proceselor de punere în aplicare ar trebui să aibă loc în mod permanent. Deoarece sistemele IA evoluează în mod continuu și acționează într-un mediu dinamic, realizarea unei IA fiabile este un proces continuu, ilustrat în figura 3 de mai jos.

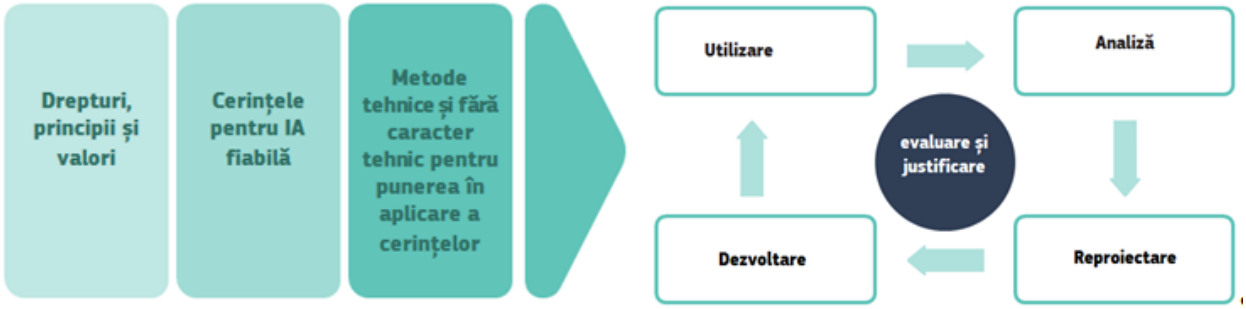

Figura 3: Asigurarea IA fiabile pe toată durata ciclului de viață al sistemului

(93) Următoarele metode pot fi considerate fie complementare, fie alternative între ele, deoarece cerințe diferite - și sensibilități diferite - pot necesita metode de punere în aplicare diferite. Această prezentare generală nu este menită să fie cuprinzătoare sau exhaustivă și nici obligatorie. Mai degrabă, are scopul de a oferi o listă de metode sugerate care pot contribui la punerea în aplicare a IA fiabile.

1. <u>Metodele tehnice</u>

(94) Această secțiune descrie metodele tehnice necesare pentru a asigura faptul că IA fiabilă poate fi integrată în fazele de proiectare, dezvoltare și utilizare a unui sistem IA. Metodele enumerate mai jos variază în ceea ce privește nivelul de maturitate.[52]

- *Arhitecturile pentru IA fiabilă*

(95) Cerințele pentru IA fiabilă ar trebui să fie „traduse” în proceduri și/sau constrângeri privind procedurile, care ar trebui să fie ancorate în arhitectura sistemului IA. Acest lucru s-ar putea realiza printr-o serie de norme aflate pe „lista albă” (comportamente sau stări) pe care sistemul ar trebui să le respecte în permanență, de restricții aflate pe „lista neagră” privind comportamentele sau stările pe care sistemul nu ar trebui să le încalce niciodată, precum și de amestecuri ale acestora sau garanții demonstrabile mai complexe privind

[50] A se vedea, de asemenea, Avizul Agenției pentru Drepturi Fundamentale a Uniunii Europene privind „Îmbunătățirea accesului la măsurile de reparare în domeniul afacerilor și al drepturilor omului la nivelul UE (2017), https://fra.europa.eu/en/opinion/2017/business-human-rights.

[51] Acest lucru implică, de exemplu, justificarea alegerilor făcute în ceea ce privește proiectarea, dezvoltarea și implementarea sistemului pentru a integra cerințele menționate anterior.

[52] Deși unele dintre aceste metode sunt deja disponibile în prezent, altele necesită în continuare mai multă cercetare. Domeniile în care este necesară o cercetare suplimentară vor trebui, de asemenea, să ofere informații pentru al doilea document al AI HLEG, și anume Recomandările în materie de politică și de investiții.

comportamentul sistemului. Monitorizarea respectării de către sistem a acestor restricții în timpul operațiunilor trebuie să se realizeze printr-un proces separat.

(96) Sistemele IA cu capacități de învățare, care își pot adapta comportamentul în mod dinamic, pot fi înțelese ca un sistem nedeterminist care poate prezenta un comportament neașteptat. Adesea, acestea sunt luate în considerare prin prisma teoretică a unui ciclu de „detectare-planificare-acționare". În vederea adaptării acestei arhitecturi pentru a asigura IA fiabilă este necesară integrarea cerințelor în toate cele trei etape ale ciclului: (i) în etapa de „detectare", sistemul ar trebui să fie dezvoltat astfel încât să recunoască toate elementele de mediu necesare pentru a asigura respectarea cerințelor; (ii) în etapa de „planificare", sistemul ar trebui să ia în considerare doar planurile care respectă cerințele; (iii) în etapa de „acționare", acțiunile sistemului ar trebui să fie restricționate la comportamentele care realizează cerințele.

(97) Arhitectura schițată mai sus este generală și nu face decât să prezinte o descriere imperfectă pentru majoritatea sistemelor IA. Totuși, aceasta oferă puncte de ancorare pentru constrângerile și politicile care ar trebui să fie reflectate în module specifice pentru a rezulta într-un sistem global care să fie fiabil și perceput ca atare.

- *Etica și statul de drept din faza de proiectare (X din faza de proiectare)*

(98) Metodele de asigurare a valorilor din faza de proiectare asigură legături exacte și explicite între principiile abstracte pe care sistemul trebuie să le respecte și deciziile specifice de punere în aplicare. Ideea că respectarea normelor poate fi pusă în aplicare în proiectarea sistemului IA este esențială pentru această metodă. Întreprinderile sunt responsabile pentru a identifica impactul sistemelor IA proprii încă de la bun început, precum și regulile pe care sistemul IA ar trebui să le respecte pentru a preveni impactul negativ. Se utilizează deja la scară largă diferite concepte „din faza de proiectare", de exemplu, *protejarea vieții private din faza de proiectare* și *securitatea din faza de proiectare*. Astfel cum se indică mai sus, pentru a câștiga încrederea, IA trebuie să fie sigură în ceea ce privește procesele, datele și rezultatele sale și ar trebui să fie proiectată astfel încât să reziste în fața datelor și atacurilor adversare. Aceasta ar trebui să pună în aplicare un mecanism pentru o închidere de siguranță și să permită reluarea funcționării după o închidere forțată (de exemplu, un atac).

- *Metodele de explicare*

(99) Pentru ca un sistem să fie fiabil, trebuie să putem înțelege de ce acesta s-a comportat într-un anumit mod și de ce a furnizat o anumită interpretare. Un domeniu de cercetare separat, IA explicabilă (XIA) încearcă să abordeze acest aspect pentru a înțelege mai bine mecanismele care stau la baza sistemului și pentru a găsi soluții. În prezent, aceasta este în continuare o provocare deschisă pentru sistemele IA bazate pe rețele de tip neural. Procesele de formare cu rețele de tip neural pot determina parametri de rețea stabiliți la valori numerice care sunt dificil de corelat cu rezultatele. În plus, micile modificările ale valorilor datelor ar putea conduce uneori la modificări majore ale interpretării, determinând, de exemplu, sistemul să confunde un autobuz școlar cu un struț. Această vulnerabilitate poate fi exploatată inclusiv în cazul atacurilor asupra sistemului. Metodele care implică cercetarea XIA sunt esențiale nu numai pentru a explica utilizatorilor comportamentul sistemului, ci și pentru a implementa o tehnologie fiabilă.

- *Testarea și validarea*

(100) Din cauza naturii nedeterministe și specifice contextului a sistemelor IA, testarea tradițională nu este suficientă. Disfuncționalitățile conceptelor și ale reprezentărilor utilizate de sistem pot apărea doar atunci când un program este aplicat unor date suficient de realiste. În consecință, pentru a verifica și a valida prelucrarea datelor, modelul de bază trebuie monitorizat cu atenție, atât pe parcursul formării, cât și al implementării, în ceea ce privește stabilitatea, soliditatea și funcționarea acestuia în cadrul unor limite bine înțelese și previzibile. Trebuie să se asigure faptul că rezultatul procesului de planificare este coerent cu intrările și că deciziile sunt adoptate într-o manieră care să permită validarea procesului de bază.

(101) Testarea și validarea sistemului ar trebui să aibă loc cât mai curând posibil, asigurând faptul că sistemul are comportamentul dorit, pe toată durata ciclului său de viață și în special după implementare. Acestea ar trebui să includă toate componentele unui sistem IA, inclusiv datele, modelele formate în prealabil, mediile și comportamentul sistemului în ansamblu. Testarea și validarea ar trebui să fie concepute și realizate de un grup de persoane cât mai divers posibil. Ar trebui să se elaboreze indicatori multipli care să acopere categoriile care sunt testate pentru diferite perspective. Poate fi luată în considerare testarea adversară de către „echipe roșii" (red teams) de încredere și diverse, care încearcă în mod deliberat să „spargă" sistemul pentru a descoperi vulnerabilitățile, precum și prin programe „Bug Bounty" care să-i stimuleze pe cei din exterior să detecteze și să raporteze în mod responsabil erorile și deficiențele sistemului. În final, trebuie să se asigure faptul că ieșirile sau acțiunile sunt coerente cu rezultatele proceselor precedente, comparându-le cu politicile definite anterior pentru a asigura faptul că acestea nu sunt încălcate.

- *Calitatea indicatorilor de servicii*

(102) Se poate defini o calitate adecvată a indicatorilor de servicii pentru sisteme IA pentru a asigura faptul că există o înțelegere de bază a faptului dacă acestea au fost testate și dezvoltate ținând cont de considerente de securitate și siguranță. Acești indicatori ar putea include măsuri de evaluare a testării și a formării algoritmilor, precum și indicatori software tradiționali ai funcționalității; performanței; ușurinței în utilizare; fiabilității; securității; și întreținerii.

2. Metodele fără caracter tehnic

(103) Această secțiune descrie o varietate de metode fără caracter tehnic care pot avea un rol important în ceea ce privește asigurarea și menținerea IA fiabile. Și aceste metode ar trebui să fie evaluate **în mod permanent**.

- *Regulamentul*

(104) Astfel cum s-a menționat anterior, există deja în prezent un regulament pentru a sprijini fiabilitatea IA - a se vedea legislația privind siguranța produselor și cadrele privind răspunderea. În măsura în care considerăm că poate fi necesară revizuirea, adaptarea sau introducerea regulamentului, atât ca garanție, cât și ca motor, acest lucru va fi abordat în al doilea document al nostru, care include Recomandări în materie de politică și de investiții în domeniul IA.

- *Codurile de conduită*

(105) Organizațiile și părțile interesate pot adera la orientări și își pot adapta carta de responsabilitate corporativă, indicatorii-cheie de performanță, codurile de conduită sau documentele de politică internă pentru a adăuga eforturile pentru realizarea IA fiabile. O organizație care lucrează la un sistem IA poate, la modul general, să își documenteze intențiile, precum și să includă standarde ale anumitor valori dorite, cum ar fi drepturile fundamentale, transparența și evitarea daunelor.

- *Standardizarea*

(106) Standardele, de exemplu pentru proiectare, producție și practicile comerciale, pot funcționa ca un sistem de management al calității pentru utilizatorii IA, consumatori, organizații, institutele de cercetare și guverne, oferind capacitatea de a recunoaște și a încuraja conduita etică prin deciziile lor de achiziție. În afară de standardele convenționale, există abordări de coreglementare: sistemele de acreditare, codurile deontologice sau standardele pentru proiectare conformă cu drepturile fundamentale. Câteva exemple actuale includ standardele ISO sau seria de standarde IEEE P7000, însă, în viitor, ar putea fi adecvată o posibilă etichetă „IA fiabilă", care că confirme, prin trimitere la standardele tehnice specifice, faptul că sistemul, de exemplu, aderă la principiile privind siguranța, soliditatea tehnică și explicabilitatea.

- *Certificare*

(107) Deoarece nu se poate preconiza că fiecare persoană poate să înțeleagă pe deplin funcționarea și efectele

sistemelor IA, pot fi luate în considerare organizațiile care atestă publicului larg faptul că un sistem IA este transparent, responsabil și echitabil[53]. Aceste certificări ar aplica standardele elaborate pentru diferite domenii de aplicare și tehnici IA, aliniate în mod adecvat la standardele industriale și sociale ale contextului diferit. Cu toate acestea, certificarea nu poate înlocui niciodată responsabilitatea. Prin urmare, aceasta ar trebui să fie completată de cadre de responsabilitate, inclusiv de declarații de declinare a responsabilității, precum și de mecanisme de revizuire și remediere[54].

- *Responsabilitatea prin intermediul cadrelor de guvernanță*

(108) Organizațiile ar trebui să instituie cadre de guvernanță, atât interne, cât și externe, care să asigure responsabilitatea pentru dimensiunile etice ale deciziilor asociate cu dezvoltarea, implementarea și utilizarea IA. Acest lucru poate include, de exemplu, numirea unei persoane responsabile pentru problemele etice legate de IA sau un comitet sau consiliu de etică intern/extern. Printre rolurile posibile ale acestei persoane sau ale acestui comitet sau consiliu se numără asigurarea controlului și consultanța. Astfel cum se prevede mai sus, specificațiile și/sau organismele de certificare pot, de asemenea, să joace un rol în acest scop. Ar trebui să se asigure canale de comunicare cu grupurile de supraveghere din industrie și/sau publice, prin care să se facă schimb de bune practici, să se discute dilemele sau să se raporteze problemele emergente care prezintă preocupări etice. Aceste mecanisme pot completa, dar nu pot înlocui supravegherea legală (de exemplu, sub forma numirii unui responsabil cu protecția datelor sau măsuri echivalente, impuse prin lege conform legislației în domeniul protecției datelor).

- *Educația și sensibilizarea cu privire la promovarea unei mentalități etice*

(109) IA fiabilă încurajează participarea informată a tuturor părților interesate. Comunicarea, educația și formarea joacă un rol important, atât pentru a asigura faptul că se cunoaște pe scară largă potențialul impact al sistemelor IA, cât și pentru a le face cunoscut oamenilor că pot participa la modelarea dezvoltării societății. Acest lucru include toate părțile interesate, de exemplu cele implicate în realizarea produselor (proiectanții și dezvoltatorii), utilizatorii (întreprinderile sau persoanele fizice) și alte grupuri afectate (cele care este posibil să nu achiziționeze sau utilizeze un sistem IA, dar pentru care deciziile sunt luate de un sistem IA, precum și societatea în general). În cadrul societății ar trebui să fie promovată alfabetizarea de bază în domeniul IA. O condiție prealabilă pentru educarea publicului este aceea de a asigura competențele adecvate și formarea specialiștilor în etică în acest domeniu.

- *Participarea părților interesate și dialogul social*

(110) Beneficiile IA sunt numeroase iar Europa trebuie să asigure faptul că acestea sunt disponibile pentru toți. În acest scop, sunt necesare o discuție deschisă și implicarea partenerilor sociali, a părților interesate, precum și a publicului larg. Numeroase organizații se bazează deja pe comitete de părți interesate pentru a discuta despre utilizarea sistemelor IA și analiza de date. Aceste comitete includ diverși membri, cum ar fi juriști, experți tehnici, specialiști în etică, reprezentanți ai consumatorilor și lucrători. Urmărirea în mod activ a participării și a dialogului cu privire la utilizarea și impactul sistemelor IA sprijină evaluarea rezultatelor și a abordărilor și poate fi extrem de utilă în cazuri complexe.

- *Diversitatea și echipe de proiectare favorabile incluziunii*

(111) Diversitatea și incluziunea joacă un rol esențial în dezvoltarea sistemelor IA care vor fi utilizate în lumea reală. Deoarece sistemele IA efectuează mai multe sarcini pe cont propriu, este esențial ca echipele care proiectează, dezvoltă, testează și întrețin, implementează și/sau achiziționează aceste sisteme să reflecte diversitatea

[53] Astfel cum se susține, de exemplu, în IEEE Ethically Aligned Design Initiative (Inițiativa de proiectare aliniată din punct de vedere etic a IEEE): https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.

[54] Pentru mai multe informații privind limitările certificării, a se vedea: https://ainowinstitute.org/AI_Now_2018_Report.pdf.

utilizatorilor și a societății în general. Acest lucru contribuie la obiectivitatea și luarea în considerare a diferitelor perspective, nevoi și obiective. În mod ideal, echipele nu sunt diverse doar în ceea ce privește genul, cultura, vârsta, ci și în ceea ce privește contextele profesionale și gama de competențe.

Orientări esențiale rezultate din capitolul II:

- ✓ Asigurarea faptului că întregul ciclu de viață al sistemului IA îndeplinește cerințele pentru o IA fiabilă: (1) factorul uman și supravegherea umană, (2) soliditatea tehnică și siguranța, (3) viața privată și guvernanța datelor, (4) transparența, (5) diversitatea, nediscriminarea și echitatea, (6) bunăstarea socială și ecologică și (7) responsabilitatea.
- ✓ Analizarea metodelor tehnice și fără caracter tehnic pentru a se asigura punerea în aplicare a acestor cerințe.
- ✓ Stimularea cercetării și a inovării pentru a contribui la evaluarea sistemelor IA și a sprijini îndeplinirea cerințelor; diseminarea rezultatelor și a întrebărilor deschise pentru publicul larg și formarea în mod sistematic a unei noi generații de experți în etica IA.
- ✓ Comunicarea, într-un mod clar și proactiv, de informații părților interesate cu privire la capacitățile și limitările sistemului IA, pentru a le permite să aibă așteptări realiste, precum și informații 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 a posibilității de auditare a sistemelor IA, în special în contexte și situații critice.
- ✓ Implicarea părților interesate pe toată durata ciclului de viață al sistemului IA. Stimularea formării și a educației, astfel încât toate părțile interesate să fie informate în ceea ce privește IA fiabilă și să fie formate în acest sens.
- ✓ Conștientizarea faptului că ar putea exista tensiuni fundamentale între diferitele principii și cerințe. Identificarea, evaluarea, documentarea și comunicarea în mod continuu a compromisurilor și a soluțiilor găsite în acest sens.

III. Capitolul III: evaluarea IA fiabile

(112) Pe baza cerințelor-cheie din capitolul II, acest capitol stabilește o **Listă** neexhaustivă **de evaluare pentru o IA fiabilă** (versiunea pilot) pentru **operaționalizarea IA fiabile**. Aceasta se aplică în special sistemelor IA care interacționează în mod direct cu utilizatorii și se adresează în principal dezvoltatorilor și organizațiilor de implementare a sistemelor IA (indiferent dacă acestea sunt dezvoltate intern sau achiziționate de la terțe părți). Această listă de evaluare nu abordează operaționalizarea primei componente a IA fiabile (IA legală). Respectarea acestei liste de evaluare nu reprezintă o dovadă de conformitate legală și nici nu este menită să ofere orientări pentru a asigura respectarea legii aplicabile. Având în vedere caracterul specific al sistemelor IA în funcție de aplicație, lista de evaluare va trebui să fie adaptată la utilizările și contextele specifice în care funcționează sistemele. În plus, acest capitol oferă o recomandare generală cu privire la modul de punere în aplicare a Listei de evaluare pentru o IA fiabilă printr-o structură de guvernanță care să cuprindă atât nivelul operațional, cât și cel al conducerii.

(113) Lista de evaluare și structura de guvernanță vor fi elaborate în strânsă colaborare cu părțile interesate din sectoarele public și privat. Acest proces se va desfășura ca proces „pilot”, permițând un feedback amplu în urma a două procese paralele:

a. un proces calitativ care asigură reprezentabilitatea și în cadrul căruia un număr mic de întreprinderi, organizații și instituții (din diferite sectoare și de dimensiuni diferite) se vor angaja să testeze lista de evaluare și structura de guvernanță în practică și să ofere feedback aprofundat;

b. un proces cantitativ în care toate părțile interesate se pot angaja să testeze lista de evaluare și să ofere feedback prin intermediul unei consultări deschise.

(114) După etapa pilot, vom integra rezultatele obținute din procesul de feedback în lista de evaluare și vom elabora o versiune revizuită la începutul anului 2020. Scopul este realizarea unui cadru care să poată fi utilizat la nivel orizontal în toate aplicațiile și, așadar, să se ofere o bază pentru asigurarea IA fiabile în toate domeniile. După stabilirea acestei baze, ar putea fi elaborat un cadru sectorial sau specific aplicațiilor.

Guvernanța

(115) Întreprinderile, organizațiile și instituțiile pot dori să ia în considerare modul în care Lista de evaluare pentru o IA fiabilă poate fi pusă în aplicare în organizația lor. Acest lucru se poate efectua prin includerea procesului de evaluare în mecanismele de guvernanță existente sau prin punerea în aplicare a unor procese noi. Această alegere va depinde de structura internă a organizației, precum și de dimensiunea și resursele sale disponibile.

(116) Studiile[55] arată că atenția conducerii la cel mai înalt nivel este esențială pentru realizarea modificării. Aceasta demonstrează, de asemenea, că implicarea tuturor părților interesate dintr-o întreprindere, organizație sau instituție stimulează acceptarea și relevanța introducerii oricăror procese noi (indiferent dacă sunt procese tehnologice sau nu)[56]. Prin urmare, recomandăm punerea în aplicare a unui proces care să cuprindă atât implicarea nivelului operațional, cât și cea a conducerii de nivel superior.

[55] https://www.mckinsey.com/business-functions/operations/our-insights/secrets-of-successful-change-implementation

[56] A se vedea, de exemplu, A. Bryson, E. Barth și H. Dale-Olsen, *The Effects of Organisational change on worker well-being and the moderating role of trade unions* (Efectele modificărilor la nivel organizatoric asupra bunăstării lucrătorilor și rolul de moderator al sindicatelor), *ILRReview*, 66(4), iulie 2013; Jirjahn, U. și Smith, S.C. (2006). „*What Factors Lead Management to Support or Oppose Employee Participation - With and Without Works Councils? Hypotheses and Evidence from Germany's Industrial Relations*" (Ce factori determină conducerea să sprijine sau să se opună participării angajaților - cu sau fără comitete de întreprindere? Ipoteze și dovezi din relațiile industriale ale Germaniei), 45(4), 650–680; Michie, J. și Sheehan, M. (2003). „*Labour market deregulation, «flexibility» and innovation*" (Dereglementarea pieței forței de muncă, „flexibilitate" și inovare), *Cambridge Journal of Economics*, 27(1), 123–143.

Nivel	Roluri relevante (în funcție de organizație)
Conducere și consiliu	Conducerea de nivel superior discută și evaluează dezvoltarea, implementarea sau achiziționarea IA și servește drept comitet de escaladare pentru evaluarea tuturor inovațiilor și utilizărilor IA atunci când se detectează preocupări majore. Aceasta implică toate părțile afectate de posibila introducere a sistemelor IA (de exemplu, lucrătorii) și reprezentanții acestora pe tot parcursul procesului, prin proceduri de informare, consultare și participare.
Departamentul pentru conformitate/Departamentul juridic/Departamentul de responsabilitate corporativă	Departamentul de responsabilitate monitorizează utilizarea listei de evaluare și evoluția necesară a acesteia pentru a face față modificărilor tehnologice sau în materie de reglementare. Acesta actualizează standardele sau politicile interne cu privire la sistemele IA și asigură faptul că utilizarea sistemelor respectă cadrul juridic și de reglementare actual, precum și valorile organizației.
Departamentul de dezvoltare a produselor și serviciilor sau un departament echivalent	Departamentul de dezvoltare a produselor și serviciilor utilizează lista de evaluare pentru a evalua produsele și serviciile bazate pe IA și înregistrează toate rezultatele. Aceste rezultate sunt discutate la nivelul conducerii, care, în cele din urmă, aprobă aplicațiile noi sau revizuite bazate pe IA.
Asigurarea calității	Departamentul de asigurare a calității (sau un departament echivalent) asigură și verifică rezultatele listei de evaluare și acționează pentru a transmite problema la un nivel superior, dacă rezultatul nu este satisfăcător sau dacă sunt detectate rezultate neprevăzute.
Resurse umane	Departamentul de resurse umane asigură combinația adecvată de competențe și diversitatea profilurilor pentru dezvoltatorii de sisteme IA. Acesta asigură faptul că, în interiorul organizației, se furnizează nivelul adecvat de formare privind IA fiabilă.
Achiziții	Departamentul de achiziții asigură faptul că procesul de achiziționare a produselor sau serviciilor bazate pe IA include o verificare a IA fiabile.
Operațiuni zilnice	Dezvoltatorii și managerii de proiect includ lista de evaluare în activitatea lor zilnică și documentează rezultatele și efectele evaluării.

Utilizarea listei de evaluare pentru o IA fiabilă

(117) Atunci când se utilizează lista de evaluare în practică, recomandăm să se acorde atenție nu numai domeniilor de interes, ci și întrebărilor la care nu se poate răspunde (cu ușurință). O problemă potențială ar putea fi lipsa de diversitate a aptitudinilor și a competențelor în cadrul echipei care dezvoltă și testează sistemul IA și, prin urmare, ar putea fi necesară implicarea altor părți interesate din interiorul sau din afara organizației. Se recomandă cu fermitate să se înregistreze toate rezultatele, atât din punct de vedere tehnic, cât și al conducerii, asigurând faptul că soluționarea problemelor poate fi înțeleasă la toate nivelurile din structura de guvernanță.

(118) Această listă de evaluare are ca scop să ofere orientări practicienilor din domeniul IA pentru a dezvolta, implementa și utiliza IA fiabilă. Evaluarea ar trebui să fie adaptată în mod proporțional la utilizarea specifică. În etapa pilot, pot fi dezvăluite domenii sensibile specifice, iar necesitatea unor specificații suplimentare în astfel de cazuri va fi evaluată în etapa următoare. Deși lista de evaluare nu oferă

răspunsuri concrete care să abordeze întrebările adresate, aceasta încurajează reflecția cu privire la măsurile care pot contribui la asigurarea fiabilității sistemelor IA, precum și la măsurile potențiale care ar trebui adoptate în această privință.

Relația cu legislația și procesele existente

(119) De asemenea, este important ca părțile implicate în dezvoltarea, implementarea și utilizarea IA să recunoască faptul că există diverse legi care impun procese specifice și interzic rezultate specifice, care se pot suprapune sau pot coincide cu unele dintre măsurile enumerate în lista de evaluare. De exemplu, legislația în domeniul protecției datelor stabilește o serie de cerințe legale care trebuie să fie îndeplinite de părțile implicate în ceea ce privește colectarea și prelucrarea datelor cu caracter personal. Cu toate acestea, deoarece IA fiabilă necesită, de asemenea, prelucrarea etică a informațiilor, procedurile și politicile interne menite să asigure respectarea legislației în domeniul protecției datelor ar putea contribui la facilitarea prelucrării etice a informațiilor și, prin urmare, pot completa procesele existente. Totuși, respectarea acestei liste de evaluare *nu* reprezintă o dovadă de conformitate legală și nici nu este menită să ofere orientări pentru a asigura respectarea legislației aplicabile. Această listă de evaluare are, mai degrabă, scopul de a oferi destinatarilor o serie de întrebări specifice, urmărind să asigure faptul că abordarea dezvoltării și a implementării IA de către aceștia este orientată către IA fiabilă și urmărește să o asigure.

(120) În mod similar, numeroși practicieni în domeniul IA au deja instrumente de evaluare și procese de dezvoltare software existente pentru a asigura inclusiv conformitatea cu standardele fără caracter juridic. Evaluarea de mai jos nu ar trebui să fie efectuată neapărat ca un exercițiu de sine stătător, ci poate fi integrată în aceste practici existente.

LISTA DE EVALUARE PENTRU O IA FIABILĂ (VERSIUNEA PILOT)

1. Factorul uman și supravegherea umană

Drepturile fundamentale:

- ✓ În cazurile de utilizare în care poate exista un impact negativ asupra drepturilor fundamentale, ați efectuat o evaluare a impactului asupra drepturilor fundamentale? Ați identificat și ați documentat potențialele compromisuri făcute între diferitele principii și drepturi?
- ✓ Sistemul IA interacționează cu luarea deciziilor de către utilizatorii finali umani (de exemplu, acțiunile sau deciziile recomandate care trebuie adoptate, prezentarea opțiunilor)?
  - În aceste cazuri, există riscul ca sistemul IA să afecteze autonomia oamenilor prin faptul că afectează procesul decizional al utilizatorului final în mod neintenționat?
  - Ați examinat dacă sistemul IA ar trebui să comunice utilizatorilor faptul că o decizie, un conținut, o recomandare sau o realizare este rezultatul unei decizii algoritmice?
  - În cazul în care sistemul IA este prevăzut cu un robot pentru chat sau cu un sistem de conversație, utilizatorii finali umani sunt informați cu privire la faptul că interacționează cu un agent neuman?

Factorul uman:

- ✓ În cazul în care sistemul IA este pus în aplicare în procesul de muncă, ați analizat repartizarea

sarcinilor între sistemul IA și lucrătorii umani astfel încât să se asigure interacțiuni semnificative, precum și pentru o supraveghere și un control adecvat de către om?

- Sistemul IA consolidează sau sporește capacitățile umane?
- Ați luat măsuri de salvgardare pentru a preveni încrederea sau importanța excesivă acordată sistemului IA în procesele de lucru?

Supravegherea umană:

✓ Ați examinat care ar fi nivelul adecvat de control uman pentru un anumit sistem IA și o anumită utilizare?

- Puteți descrie nivelul de control uman sau de implicare a omului, dacă este cazul? Cine este „persoana care deține controlul" și care sunt momentele sau instrumentele de intervenție umană?
- Ați instituit mecanisme și măsuri pentru a asigura acest control potențial sau această supraveghere potențială umană sau pentru a asigura faptul că deciziile sunt adoptate sub responsabilitatea globală a oamenilor?
- Ați adoptat măsuri care să permită auditarea și să remedieze problemele legate de guvernarea autonomiei IA?

✓ În cazul unui sistem IA autonom sau cu capacitate de auto-învățare sau al utilizării unui astfel de sistem, ați instituit mai multe mecanisme specifice de control și supraveghere?

- Ce tip de mecanisme de detectare și de răspuns ați instituit pentru a evalua dacă ar putea apărea probleme?
- Ați asigurat un „buton de oprire" sau o procedură pentru a întrerupe o operațiune, dacă este necesar? Această procedură întrerupe procesul în întregime, parțial sau deleagă controlul unui om?

2. <u>Soliditatea tehnică și siguranța</u>

Reziliența la atacuri și securitatea:

✓ Ați evaluat potențialele forme de atac la care sistemul IA ar putea fi vulnerabil?

- În special, ați luat în considerare diferitele tipuri și naturi ale vulnerabilităților, cum ar fi poluarea datelor, infrastructura fizică, atacurile cibernetice?

✓ Ați instituit măsuri sau sisteme care să asigure integritatea și reziliența sistemului IA la atacurile potențiale?

✓ Ați evaluat modul în care se comportă sistemul dumneavoastră în situații și medii neașteptate?

✓ Ați examinat dacă și în ce măsură sistemul dumneavoastră ar putea să aibă dublă utilizare? În caz afirmativ, ați luat măsuri preventive adecvate împotriva acestei situații (inclusiv, de exemplu,

nepublicarea cercetării sau neimplementarea sistemului)?

Soluțiile de siguranță și siguranța generală:

- ✓ Ați asigurat faptul că sistemul dumneavoastră dispune de suficiente soluții de siguranță în cazul în care se confruntă cu atacuri adversare sau cu alte situații neașteptate (de exemplu, proceduri de transfer tehnic sau solicitarea unei acțiuni din partea unui operator uman înainte de operare)?
- ✓ Ați examinat nivelul de risc pe care îl prezintă sistemul IA în acest caz de utilizare specifică?
  - Ați instituit un proces pentru a măsura și a evalua riscurile și siguranța?
  - Ați furnizat informațiile necesare în cazul unui risc pentru integritatea fizică a oamenilor?
  - Ați luat în considerare o poliță de asigurare pentru potențialele daune provocate de sistemul IA?
  - Ați identificat potențialele riscuri la adresa siguranței pe care le prezintă (alte) utilizări previzibile ale tehnologiei, inclusiv utilizarea greșită accidentală sau rău intenționată a acesteia? Există un plan de atenuare sau de gestionare a acestor riscuri?
- ✓ Ați evaluat dacă există probabilitatea ca sistemul IA să cauzeze prejudicii sau daune utilizatorilor sau terților părți? În caz afirmativ, ați evaluat probabilitatea, prejudiciile potențiale, publicul afectat și gravitatea?
  - În cazul în care există riscul ca sistemul IA să cauzeze prejudicii, ați luat în considerarea răspunderea și normele de protecție a consumatorilor și cum ați luat în considerare aceste aspecte?
  - Ați luat în considerare impactul potențial sau riscul în materie de siguranță pentru mediu sau animale?
  - Analiza riscurilor pe care o efectuați ia în considerare dacă problemele de securitate sau de rețea (de exemplu, pericolele cibernetice) prezintă riscuri în materie de siguranță sau cauzează prejudicii din cauza comportamentului neintenționat al sistemului IA?
- ✓ Ați estimat impactul probabil al unei disfuncționalități a sistemului IA care determină furnizarea unor rezultate incorecte, care face ca sistemul dumneavoastră să fie indisponibil sau să furnizeze rezultate inacceptabile din punct de vedere social (de exemplu, practici discriminatorii)?
  - Ați definit praguri și o structură de guvernanță pentru scenariile de mai sus în vederea declanșării planurilor alternative/soluțiilor de siguranță?
  - Ați definit și ați testat soluțiile de siguranță?

Acuratețe

- ✓ Ați evaluat ce nivel și ce definiție a acurateței ar fi necesare în contextul sistemului IA și al utilizării acestuia?
  - Ați evaluat modul în care este măsurată și asigurată acuratețea?
  - Ați instituit măsuri pentru a asigura faptul că datele utilizate sunt complete și actualizate?
  - Ați instituit măsuri pentru a evalua dacă sunt necesare date suplimentare, de exemplu pentru a

îmbunătăți acuratețea sau pentru a elimina judecățile părtinitoare?

- ✓ Ați evaluat daunele potențiale dacă sistemul IA face previziuni inexacte?
- ✓ Ați instituit modalități de a măsura dacă sistemul dumneavoastră face un număr inacceptabil de previziuni inexacte?
- ✓ Dacă se fac previziuni inexacte, ați instituit o serie de măsuri pentru soluționarea acestei probleme?

Fiabilitatea și reproductibilitatea:

- ✓ Ați instituit o strategie pentru a monitoriza și a testa dacă sistemul IA realizează obiectivele, scopurile și corespunde aplicațiilor avute în vedere?
  - Ați testat dacă trebuie să se țină seama de contexte specifice sau de anumite condiții pentru a asigura reproductibilitatea?
  - Ați instituit procese sau metode de verificare pentru a măsura și a asigura diferitele aspecte ale fiabilității și reproductibilității?
  - Ați instituit procese pentru a descrie disfuncționalitățile sistemului IA în anumite contexte?
  - Ați documentat în mod clar și ați operaționalizat aceste procese pentru a testa și a verifica fiabilitatea sistemelor IA?

    Ați instituit un mecanism sau o comunicare pentru a garanta utilizatorilor (finali) fiabilitatea sistemului IA?

3. <u>Viața privată și guvernanța datelor</u>

Respectarea vieții private și protecția datelor:

- ✓ În funcție de utilizare, ați instituit mecanisme care să permită altora să semnaleze problemele legate de viața privată sau de protecția datelor cu privire la procesele sistemului IA de colectare de date (pentru formare, precum și pentru funcționare) și de prelucrare a datelor?
- ✓ Ați evaluat tipul și domeniul de aplicare al datelor din seturile dumneavoastră de date (de exemplu, dacă acestea conțin date cu caracter personal)?
- ✓ Ați luat în considerare modalități de dezvoltare a sistemului IA sau de formare a modelului cu sau fără utilizarea minimă a datelor potențial sensibile sau cu caracter personal?
- ✓ Ați integrat mecanisme de notificare și control cu privire la datele cu caracter personal în funcție de utilizare (cum ar fi consimțământul valabil și posibilitatea de revocare, dacă este cazul)?
- ✓ Ați adoptat măsuri pentru a consolida respectarea vieții private, de exemplu, prin criptare, anonimizare și agregare?
- ✓ În cazul în care există un responsabil cu protecția datelor, ați implicat această persoană în proces într-un stadiu incipient?

Calitatea și integritatea datelor:

- ✓ V-ați aliniat sistemul la standardele potențiale relevante (de exemplu, ISO, IEEE) sau la protocoalele

adoptate pe scară largă pentru gestionarea și guvernanța zilnică a datelor dumneavoastră?

- ✓ Ați instituit mecanisme de supraveghere pentru colectarea, stocarea, prelucrarea și utilizarea datelor?
- ✓ Ați evaluat măsura în care dețineți controlul asupra calității surselor externe de date utilizate?
- ✓ Ați instituit procese pentru a asigura calitatea și integritatea datelor dumneavoastră? Ați luat în considerare alte procese? Cum verificați dacă seturile dumneavoastră de date nu au fost compromise sau atacate de hackeri?

Accesul la date:

- ✓ Ce protocoale, procese și proceduri au fost urmate pentru a gestiona și a asigura o guvernanță adecvată a datelor?
  - Ați evaluat cine poate accesa datele utilizatorilor și în ce circumstanțe?
  - Ați asigurat faptul că aceste persoane sunt calificate și au obligația să acceseze datele și că au competențele necesare pentru a înțelege detaliile politicii de protecție a datelor?
  - Ați asigurat un mecanism de supraveghere care să înregistreze când, unde, cum, de către cine și în ce scop sunt accesate datele?

4. <u>Transparență</u>

Trasabilitatea:

- ✓ Ați instituit măsuri care să asigure trasabilitatea? Acest aspect ar putea include documentarea cu privire la următoarele:
  - ➢ Metodele utilizate pentru proiectarea și dezvoltarea sistemului algoritmic:
    - o în cazul unui sistem IA bazat pe reguli, ar trebui să se documenteze metoda de programare sau modul în care a fost construit modelul;
    - o în cazul unui sistem IA bazat pe învățare, ar trebui să se documenteze metoda de formare a algoritmului, inclusiv ce date de intrare au fost colectate și selectate, precum și modul în care s-a realizat acest lucru.
  - ➢ Metodele utilizate pentru testarea și validarea sistemului algoritmic:
    - o în cazul unui sistem IA bazat pe reguli, ar trebui să se documenteze scenariile sau cazurile utilizate în scopul testării și al validării;
    - o în cazul unui model bazat pe învățare, ar trebui să se documenteze informațiile cu privire la datele utilizate pentru testare și validare.
  - ➢ Rezultatele sistemului algoritmic:
    - o Rezultatele sau deciziile adoptate de algoritm, precum și alte decizii potențiale care ar rezulta din diferite cazuri (de exemplu, pentru alte subgrupuri de utilizatori), ar trebui să fie documentate.

Explicabilitatea:

- ✓ Ați evaluat măsura în care pot fi înțelese deciziile și, așadar, rezultatul sistemului IA?
- ✓ Ați asigurat faptul că o explicație privind motivul pentru care un sistem a făcut o anumită alegere care a condus la un anumit rezultat poate fi ușor de înțeles de către toți utilizatorii care doresc o explicație?
- ✓ Ați evaluat în ce măsură decizia sistemului influențează procesele decizionale ale organizației?
- ✓ Ați evaluat motivul pentru care a fost implementat un anumit sistem specific într-un anumit domeniu?
- ✓ Ați evaluat modelul de afaceri referitor la acest sistem (de exemplu, modul în care acesta creează valoare pentru organizație)?
- ✓ Ați proiectat sistemul IA ținând seama de interpretabilitate de la bun început?
    - Ați cercetat și ați încercat să utilizați modelul cel mai simplu și mai ușor de interpretat posibil pentru aplicația în cauză?
    - Ați evaluat dacă puteți analiza datele dumneavoastră de formare și testare? Puteți modifica și actualiza acest lucru în timp?
    - Ați evaluat dacă aveți opțiuni după formarea și dezvoltarea modelului pentru a examina interpretabilitatea sau dacă aveți acces la fluxul de lucru intern al modelului?

Comunicarea:

- ✓ Ați comunicat utilizatorilor (finali) - printr-o declarație de declinare a responsabilității sau prin orice alte mijloace - că aceștia interacționează cu un sistem IA și nu cu un alt om? Ați indicat în mod clar că sistemul dumneavoastră este un sistem IA?
- ✓ Ați instituit mecanisme de informare a utilizatorilor cu privire la motivele și criteriile care stau la baza rezultatelor sistemului IA?
    - Acest lucru este comunicat în mod clar și inteligibil utilizatorilor vizați?
    - Ați instituit procese care țin seama de feedbackul utilizatorilor și ați utilizat acest feedback pentru a adapta sistemul?
    - Ați comunicat, de asemenea, cu privire la riscurile potențiale sau percepute, cum ar fi judecățile părtinitoare?
    - În funcție de utilizare, ați luat în considerare, de asemenea, comunicarea și transparența față de alt public, terțe părți și publicul larg?
- ✓ Ați clarificat care este scopul sistemului IA și cine sau ce poate beneficia de produs/serviciu?
    - Scenariile de utilizare pentru produs au fost specificate și comunicate în mod clar, luând în considerare inclusiv forme alternative de comunicare pentru a asigura faptul că acesta poate fi înțeles și este adecvat pentru utilizatorul căruia îi este destinat?
    - În funcție de utilizare, v-ați gândit la psihologia umană și la potențialele limitări, cum ar fi riscul de confuzie, judecățile părtinitoare privind confirmarea sau oboseala cognitivă?
- ✓ Ați comunicat în mod clar caracteristicile, limitările și deficiențele potențiale ale sistemului IA:

- în cazul dezvoltării: oricărei persoane care îl implementează într-un produs sau serviciu?
- în cazul implementării: utilizatorului final sau consumatorului?

5. <u>Diversitate, nediscriminare și echitate</u>

Evitarea judecăților părtinitoare neechitabile:

✓ Ați asigurat o strategie sau o serie de proceduri pentru a evita crearea sau consolidarea judecăților părtinitoare neechitabile în sistemul IA, atât cu privire la utilizarea datelor de intrare, cât și pentru proiectarea algoritmului?

- Ați evaluat și ați recunoscut posibilele limitări care rezultă din componența seturilor de date utilizate?
- Ați luat în considerare diversitatea și reprezentativitatea utilizatorilor din date? Ați testat sistemul pentru populații specifice sau utilizări problematice?
- Ați cercetat și ați utilizat instrumentele tehnice disponibile pentru a vă îmbunătăți înțelegerea datelor, a modelului și a performanței?
- Ați instituit procese pentru testarea și monitorizarea eventualelor judecăți părtinitoare în etapa de dezvoltare, implementare și utilizare a sistemului?

✓ În funcție de utilizare, ați asigurat un mecanism care să permită altor persoane să semnaleze problemele legate de judecățile părtinitoare, discriminare sau performanța slabă a sistemului IA?

- Ați luat în considerare măsuri clare și căi de comunicare cu privire la modul în care și persoana pentru care pot apărea aceste probleme?
- Ați luat în considerare nu numai utilizatorii (finali), ci și alte părți care ar putea fi afectate în mod indirect de sistemul IA?

✓ Ați evaluat dacă în condiții identice poate apărea vreo variabilitate a deciziei?

- În caz afirmativ, ați luat în considerare care ar putea fi cauzele posibile?
- În cazul variabilității, ați instituit un mecanism de măsurare sau de evaluare a impactului potențial al acestei variabilități asupra drepturilor fundamentale?

✓ Ați asigurat o definiție de lucru adecvată a „echității” pe care să o aplicați în procesul de proiectare a sistemelor IA?

- Definiția dumneavoastră este utilizată în mod obișnuit? Ați luat în considerare alte definiții înainte de a o alege pe aceasta?
- Ați asigurat o analiză cantitativă sau indicatori pentru a măsura și a testa definiția aplicată a echității?
- Ați instituit mecanisme care să asigure echitatea sistemelor IA? Ați luat în considerare alte mecanisme potențiale?

Accesibilitate și proiectarea universală:

- ✓ Ați asigurat faptul că sistemul IA cuprinde o gamă largă de preferințe și abilități individuale?
  - Ați evaluat dacă sistemul IA poate fi utilizat de persoanele cu nevoi speciale sau cu handicap sau de cele expuse riscului de excluziune? Cum a fost proiectat în sistem și cum este verificat acest lucru?
  - Ați asigurat faptul că informațiile cu privire la sistem sunt disponibile și pentru utilizatorii de tehnologii de asistență?
  - Ați implicat sau ați consultat această comunitate în etapa de dezvoltare a sistemului IA?
- ✓ Ați luat în considerare impactul sistemului IA asupra utilizatorilor potențiali?
  - Echipa implicată în construirea sistemului IA este reprezentativă pentru utilizatorii dumneavoastră țintă? Aceasta este reprezentativă pentru populație în general, având în vedere și alte grupuri care ar putea fi afectate în mod tangențial?
  - Ați evaluat dacă există persoane sau grupuri care ar putea fi afectate în mod disproporționat de implicațiile negative?
  - Ați primit feedback de la alte echipe sau grupuri care reprezintă medii sau experiențe diferite?

Participarea părților interesate:

- ✓ Ați luat în considerare un mecanism pentru a include participarea diferitelor părți interesate în dezvoltarea și utilizarea sistemului IA?
- ✓ Ați pregătit calea pentru introducerea sistemului IA în organizația dumneavoastră, informând și implicând în prealabil lucrătorii afectați și reprezentanții acestora?

6. <u>Bunăstare socială și ecologică</u>

IA durabilă și ecologică:

- ✓ Ați instituit mecanisme de măsurare a impactului asupra mediului pe care îl au dezvoltarea, implementarea și utilizarea sistemului IA (de exemplu, energia utilizată de centrul de date, tipul de energie utilizată de centrele de date etc.)?
- ✓ Ați asigurat măsuri de reducere a impactului asupra mediului pe care îl are ciclul de viață al sistemului IA?

Impactul social:

- ✓ În cazul în care sistemul IA interacționează cu oamenii în mod direct:
  - Ați evaluat dacă sistemul IA încurajează oamenii să dezvolte atașament sau empatie față de sistem?
  - Ați asigurat faptul că sistemul IA semnalează în mod clar faptul că interacțiunea sa socială este simulată și că nu are capacitatea de „a înțelege” și de „a avea sentimente”?
- ✓ Ați asigurat faptul că impactul social al sistemului IA este bine înțeles? De exemplu, ați evaluat dacă există riscul de pierdere a locului de muncă sau de scădere a nivelului de competențe al forței de

muncă? Ce măsuri au fost adoptate pentru a contracara aceste riscuri?

Societate și democrație:

✓ Ați evaluat impactul social mai amplu al utilizării sistemului IA dincolo de utilizatorul (final) individual, cum ar fi părțile interesate potențial afectate în mod indirect?

7. <u>Responsabilitate</u>

Posibilitatea auditării:

✓ Ați instituit mecanisme care să faciliteze posibilitatea auditării sistemului de către actori interni și/sau independenți, cum ar fi asigurarea trasabilității și înregistrarea proceselor și a rezultatelor sistemului IA?

Reducerea la minimum și raportarea impactului negativ:

✓ Ați efectuat o evaluare a riscului sau a impactului sistemului IA, care să ia în considerare diferitele părți interesate care sunt afectate în mod direct sau indirect?

✓ Ați instituit cadre de formare și educare pentru a dezvolta practici de asumare a răspunderii?

- Ce lucrători sau compartimente ale echipei sunt implicate/implicați? Acest lucru depășește etapa de dezvoltare?
- În cadrul acestor sesiuni de formare se predă și cadrul juridic potențial aplicabil sistemului IA?
- Ați luat în considerare instituirea unui „comitet de evaluare etică a IA” sau a unui mecanism similar care să discute responsabilitatea globală și practicile etice, inclusiv zonele gri potențial neclare?

✓ Pe lângă inițiativele sau cadrele interne de supraveghere a eticii și a responsabilității, există tipuri de orientări externe sau au fost instituite procese de auditare?

✓ Există procese pentru ca terțele părți (de exemplu, furnizori, consumatori, distribuitori/vânzători) sau lucrătorii să raporteze potențialele vulnerabilități, riscuri sau judecăți părtinitoare din sistemul/aplicația IA?

Documentarea compromisurilor:

✓ Ați instituit un mecanism pentru a identifica interesele relevante și valorile pe care le implică sistemul IA și potențialele compromisuri dintre acestea?

✓ Ce proces utilizați pentru a decide cu privire la aceste compromisuri? Ați asigurat faptul că decizia privind compromisurile a fost documentată?

Capacitatea de reparare:

✓ Ați instituit o serie adecvată de mecanisme care să permită măsuri de reparare în cazul apariției unor daune sau a unui impact negativ?

✓ Ați instituit mecanisme pentru a furniza informații utilizatorilor (finali)/terților părți cu privire la oportunitățile de reparare?

Invităm toate părțile interesate să testeze această listă de evaluare în practică și să ofere feedback cu privire la capacitatea de punere în aplicare, la caracterul său complet, la relevanța sa pentru aplicația sau domeniul IA specific, precum și la suprapunerea sau complementaritatea acesteia cu procesele de conformitate sau de evaluare existente. Pe baza acestui feedback, o versiune revizuită a Listei de evaluare pentru o IA fiabilă va fi propusă Comisiei la începutul anului 2020.

Orientări esențiale rezultate din capitolul III:

- ✓ Adoptarea unei **Liste de evaluare** pentru o IA fiabilă atunci când se dezvoltă, se implementează ori se utilizează IA și adaptarea acesteia în funcție de utilizarea specifică a sistemului.
- ✓ Luarea în considerare a faptului că această listă de evaluare nu va fi niciodată exhaustivă. Asigurarea faptului că IA fiabilă nu înseamnă bifarea unor căsuțe, ci identificarea în mod continuu a cerințelor, evaluarea soluțiilor și asigurarea unor rezultate îmbunătățite pe parcursul întregului ciclu de viață al sistemului IA și implicarea părților interesate în acest proces.

C. EXEMPLE DE OPORTUNITĂȚI ȘI DE PREOCUPĂRI MAJORE GENERATE DE IA

(121) Secțiunea următoare conține exemple de dezvoltare și utilizare a IA care ar trebui încurajate, precum și exemple de situații în care dezvoltarea, implementarea sau utilizarea IA pot să contravină valorilor noastre și să genereze preocupări specifice. Trebuie să existe un echilibru între ceea ce ar trebui să se facă și ceea ce se poate face cu IA și trebuie să se acorde atenția cuvenită pentru ceea ce nu ar trebui să se facă cu IA.

1. Exemple de oportunități ale IA fiabile

(122) IA fiabilă poate reprezenta o oportunitate importantă de a sprijini atenuarea provocărilor presante cu care se confruntă societatea, cum ar fi îmbătrânirea populației, inegalitatea socială tot mai crescută și poluarea mediului. De asemenea, acest potențial se reflectă la nivel global, de exemplu în obiectivele de dezvoltare durabilă ale ONU[57]. Următoarea secțiune prezintă modul în care ar trebui încurajată o strategie europeană în materie de IA care să abordeze unele dintre aceste provocări.

a. Politicile climatice și infrastructura durabilă

(123) Deși combaterea schimbărilor climatice ar trebui să constituie o prioritate pentru factorii de decizie de la nivel mondial, transformarea digitală și IA fiabilă au un mare potențial de a reduce impactul oamenilor asupra mediului și de a permite utilizarea eficientă și eficace a energiei și a resurselor naturale.[58]IA fiabilă poate, de exemplu, să fie cuplată cu volumele mari de date pentru a identifica mai exact necesitățile în materie de energie, ceea ce ar duce la o infrastructură energetică și la un consum de energie mai eficiente[59].

(124) Luând în considerare sectoare precum transportul public, sistemele IA pentru sisteme de transport inteligente[60] pot fi utilizate pentru a reduce la minimum cozile de așteptare, pentru a optimiza traseele, pentru

[57] https://sustainabledevelopment.un.org/?menu=1300

[58] O serie de proiecte ale UE vizează dezvoltarea unor rețele inteligente și a unor instalații pentru stocarea energiei, care au potențialul de a contribui la o tranziție energetică cu sprijin digital de succes, inclusiv prin soluții bazate pe IA și alte soluții digitale. Pentru a completa activitatea acelor proiecte individuale, Comisia a lansat inițiativa BRIDGE, care permite proiectelor de rețele inteligente și de stocare a energiei în curs, în cadrul Orizont 2020, să creeze o viziune comună asupra problemelor transversale: https://www.h2020-bridge.eu/.

[59] A se vedea, de exemplu, proiectul Encompass: http://www.encompass-project.eu/.

[60] Noile soluții bazate pe IA contribuie la pregătirea orașelor pentru viitorul mobilității. A se vedea, de exemplu, proiectul finanțat de UE intitulat Fabulos: https://fabulos.eu/.

a permite persoanelor cu deficiențe de vedere să fie mai independente,[61] pentru a optimiza motoarele eficiente din punct de vedere energetic și pentru a consolida, astfel, eforturile de decarbonizare și a reduce amprenta de mediu pentru o societate mai verde. În prezent, la nivel mondial, un om moare la fiecare 23 de secunde într-un accident rutier[62]. Sistemele IA ar putea contribui la reducerea semnificativă a numărului de decese, de exemplu prin timpi de reacție mai buni și o mai bună respectare a normelor[63].

b. Sănătatea și bunăstarea

(125) Tehnologiile de IA fiabilă pot fi utilizate - și sunt utilizate deja - pentru a face tratamentele mai inteligente și mai bine orientate, contribuind la prevenirea bolilor care pun în pericol viața[64]. Medicii și profesioniștii din domeniul medical pot, eventual, să efectueze o analiză mai exactă și mai detaliată a datelor medicale complexe ale unui pacient, chiar înainte ca pacientul să se îmbolnăvească și să ofere un tratament preventiv adaptat[65]. În contextul îmbătrânirii populației din Europa, IA și robotica pot fi instrumente valoroase care să asiste persoanele care asigură servicii de îngrijire și să sprijine îngrijirea persoanelor în vârstă,[66] precum și să monitorizeze starea pacienților în timp real, salvând astfel vieți[67].

(126) De asemenea, IA fiabilă poate acorda asistență la o scară mai amplă. De exemplu, aceasta poate să examineze și să identifice tendințele generale din sectorul asistenței medicale și al tratamentelor,[68] conducând la depistarea precoce a bolilor, la elaborarea mai eficientă a medicamentelor, la tratamente mai bine orientate[69] și, în final, la salvarea mai multor vieți.

c. Educația de calitate și transformarea digitală

(127) Noile schimbări tehnologice, economice și de mediu înseamnă că societatea trebuie să devină mai proactivă. Guvernele, liderii din industrie, instituțiile de învățământ și organizațiile sindicale se confruntă cu responsabilitatea de a aduce cetățenii în noua eră digitală, asigurând faptul că aceștia au competențele

[61] A se vedea, de exemplu, proiectul PRO4VIP, care face parte din strategia europeană Vision 2020 de combatere a orbirii evitabile prin prevenție, în special cauzată de îmbătrânire. Mobilitatea și orientarea au constituit unul dintre domeniile prioritare ale proiectului.

[62] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[63] Proiectul european UP-Drive, de exemplu, vizează să abordeze provocările subliniate legate de transport, asigurând contribuții care permit automatizarea treptată a vehiculelor și colaborarea dintre acestea, facilitând un sistem de transport mai sigur, mai favorabil incluziunii și mai accesibil. https://up-drive.eu/.

[64] A se vedea, de exemplu, proiectul REVOLVER (Repeated Evolution of Cancer - Evoluția repetată a cancerului): https://www.healtheuropa.eu/personalised-cancer-treatment/87958/ sau proiectul Murab, care efectuează biopsii mai exacte și care vizează diagnosticarea mai rapidă a cancerului și a altor boli: https://ec.europa.eu/digital-single-market/en/news/murab-eu-funded-project-success-story.

[65] A se vedea, de exemplu, proiectul INCITE: www.karolinska.se/en/live-incite. Acest consorțiu de achizitori de asistență medicală provoacă industria să dezvolte soluții IA inteligente și alte soluții TIC care să permită intervenții la nivelul stilului de viață în procesul perioperatoriu. Obiectivul vizează noi soluții inovatoare de e-sănătate, care pot influența pacienții într-un mod personalizat pentru a întreprinde acțiunile necesare, atât înainte, cât și după intervenția chirurgicală, în ceea ce privește stilul lor de viață, pentru a optimiza rezultatul asistenței medicale.

[66] Proiectul CARESSES finanțat de UE vizează roboții pentru îngrijirea persoanelor în vârstă, axându-se pe sensibilitatea lor culturală: aceștia își adaptează modul de acțiune și de vorbire pentru a se potrivi culturii și obiceiurilor persoanei în vârstă căreia îi oferă asistență: http://caressesrobot.org/en/project/. A se vedea, de asemenea, aplicația IA denumită Alfred, un asistent virtual care ajută persoanele în vârstă să rămână active: https://ec.europa.eu/digital-single-market/en/news/alfred-virtual-assistant-helping-older-people-stay-active. În plus, proiectul EMPATTICS (EMpowering PAtients for a BeTTer Information and improvement of the Communication Systems - Capacitarea pacienților pentru o mai bună informare și îmbunătățirea sistemelor de comunicare) va cerceta și va defini modul în care profesioniștii din domeniul asistenței medicale și pacienții utilizează tehnologiile TIC, inclusiv sistemele IA pentru a planifica intervențiile pentru pacienți și pentru a monitoriza evoluția stării lor fizice și psihice: www.empattics.eu.

[67] A se vedea, de exemplu MyHealth Avatar (www.myhealthavatar.eu), care oferă o reprezentare digitală a stării de sănătate a pacientului. Proiectul de cercetare a lansat o aplicație și o platformă online care colectează informațiile dumneavoastră digitale privind starea de sănătate pe termen lung și oferă acces la acestea. Proiectul devine, deci, un partener de sănătate pe toată durata vieții („avatar”). De asemenea, MyHealthAvatar anticipează riscul dumneavoastră de atac cerebral, diabet, boală cardiacă și hipertensiune.

[68] A se vedea, de exemplu, proiectul ENRICHME (www.enrichme.eu), care abordează scăderea progresivă a capacității cognitive la populația în vârstă. O platformă integrată pentru asistență pentru autonomie la domiciliu și un robot de servicii mobile pentru monitorizarea pe termen lung și interacțiune vor ajuta persoanele în vârstă să rămână independente și active mai mult timp.

[69] A se vedea, de exemplu, utilizarea IA de către Sophia Genetics, care folosește deducția statistică, recunoașterea modelelor și învățarea automată pentru a maximiza valoarea datelor genomice și radiomice: https://www.sophiagenetics.com/home.html.

adecvate pentru a ocupa locurile de muncă ale viitorului. Tehnologiile de IA fiabilă ar putea ajuta la preconizarea locurilor de muncă și a profesiilor care vor fi perturbate de tehnologie, a rolurilor noi care vor fi create și a competențelor care vor fi necesare. Acest lucru ar putea ajuta guvernele, organizațiile sindicale și industria la planificarea calificării și conversiei profesionale a lucrătorilor. De asemenea, acest lucru ar putea oferi cetățenilor care se tem de concediere o cale de dezvoltare într-un rol nou.

(128) În plus, IA poate fi un instrument important pentru combaterea inegalităților în domeniul educației și pentru crearea unor programe educaționale personalizate și adaptabile, care ar putea permite tuturor să dobândească noi calificări, abilități și competențe în funcție de capacitatea de a învăța a fiecăruia[70]. Aceasta ar putea spori atât viteza de învățare, cât și calitatea educației - de la școala primară la universitate.

2. Exemple de preocupări majore generate de IA

(129) Preocupările majore generate de IA apar atunci când este încălcată una dintre componentele IA fiabile. Multe dintre preocupările enumerate mai jos vor intra deja în domeniul de aplicare al cerințelor legale existente, care sunt obligatorii și, prin urmare, trebuie să fie îndeplinite. Totuși, chiar și în cazul în care s-a demonstrat îndeplinirea cerințelor legale, este posibil ca acestea să nu abordeze gama completă de preocupări etice care pot apărea. Deoarece înțelegerea de către noi a adecvării normelor și a principiilor etice evoluează în mod invariabil și se poate schimba în timp, următoarea listă neexhaustivă de preocupări poate fi scurtată, extinsă, editată sau actualizată în viitor.

a. <u>Identificarea și urmărirea persoanelor fizice cu ajutorul IA</u>

(130) IA permite identificarea și mai eficientă a persoanelor fizice atât de către entitățile publice, cât și de către cele private. Printre exemplele demne de reținut ale tehnologiei adaptabile de identificare prin IA se numără recunoașterea facială și alte metode involuntare de identificare, ce utilizează datele biometrice (și anume, detectarea minciunilor, evaluarea personalității prin intermediul micro-expresiilor și detectarea vocală automată). Uneori, identificarea persoanelor fizice este rezultatul dorit, aliniat la principiile etice (de exemplu, pentru depistarea cazurilor de fraudă, de spălare a banilor sau de finanțare a terorismului). Cu toate acestea, identificarea automată generează preocupări importante, atât de natură juridică, cât și etică, deoarece poate avea un impact neașteptat asupra multor niveluri psihologice și socioculturale. Este necesară o utilizare proporțională a tehnicilor de control în domeniul IA pentru a menține autonomia cetățenilor europeni. Pentru dezvoltarea IA fiabile va fi esențial să se determine în mod clar când, cum și dacă IA poate fi utilizată pentru identificarea automată a persoanelor fizice și să se facă diferența între identificarea unei persoane fizice și depistarea și urmărirea unei persoane fizice, precum și între supravegherea individualizată și supravegherea în masă. Aplicarea acestor tehnologii trebuie să fie prevăzută în mod clar în legislația existentă[71]. În cazul în care temeiul juridic pentru această activitate este „consimțământul", trebuie să se elaboreze mijloace practice[72] care să permită consimțământul semnificativ și verificat care trebuie acordat pentru ca persoana să fie identificată în mod automat prin tehnologii IA sau echivalente. Acest lucru se aplică și utilizării datelor cu caracter personal „anonime" care pot fi repersonalizate.

b. <u>Sistemele IA ascunse</u>

(131) Oamenii ar trebui să știe întotdeauna dacă interacționează în mod direct cu un alt om sau cu o mașină și este

[70] A se vedea, de exemplu, proiectul MaTHiSiS, menit să ofere o soluție pentru învățarea bazată pe afect într-un mediu de învățare confortabil, care cuprinde dispozitive tehnologice performante și algoritmi: http://mathisis-project.eu/ A se vedea, de asemenea, Watson Classroom de la IBM sau platforma Century Tech.

[71] În această privință, amintim articolul 6 din RGPD, care prevede, printre altele, că prelucrarea datelor este legală numai dacă există un temei juridic valabil.

[72] Astfel cum arată mecanismele actuale pentru acordarea consimțământului informat pe internet, de obicei, consumatorii își dau consimțământul fără a acorda o atenție suficientă acestui aspect. Așadar, aceste mecanisme cu greu pot fi clasificate ca fiind practice.

responsabilitatea practicienilor în domeniul IA ca acest lucru să se realizeze în mod fiabil. Prin urmare, practicienii în domeniul IA ar trebui să asigure că oamenii sunt informați cu privire la - sau pot să solicite și să valideze - faptul că interacționează cu un sistem IA (de exemplu, prin emiterea unor declarații de declinare a responsabilității clare și transparente). A se reține că există cazuri limită care complică acest aspect (de exemplu, o voce umană filtrată prin IA). Ar trebui reținut faptul că o confuzie între oameni și mașini ar putea avea consecințe multiple, cum ar fi atașamentul, influența sau reducerea valorii ființei umane.[73] Prin urmare, dezvoltarea roboților umanoizi[74] ar trebui să fie supusă unei evaluări etice atente.

c. IA a permis evaluarea cetățenilor cu încălcarea drepturilor fundamentale

(132) Societățile ar trebui să depună eforturi pentru a proteja libertatea și autonomia tuturor cetățenilor. Orice formă de evaluare a cetățenilor poate conduce la pierderea acestei autonomii și poate pune în pericol principiul nediscriminării. Evaluarea ar trebui să fie utilizată numai în cazul în care există o justificare clară și dacă măsurile sunt proporționale și echitabile. Evaluarea normativă a cetățenilor (evaluarea generală a „personalității morale" sau a „integrității etice") în *toate* aspectele și pe scară largă de către autoritățile publice sau actorii privați pune în pericol aceste valori, în special atunci când nu este utilizată în conformitate cu drepturile fundamentale și când este utilizată în mod disproporționat și fără un scop legitim delimitat și comunicat.

(133) În prezent, evaluarea cetățenilor - la scară largă sau redusă - se utilizează deja în evaluările pur descriptive sau specifice domeniului (de exemplu, sistemele școlare, e-learning și permisele de conducere). Chiar și în aceste aplicații mai restrânse, ar trebui să se pună la dispoziția cetățenilor o procedură complet transparentă, inclusiv informații cu privire la procesul, scopul și metodologia evaluării. Trebuie remarcat faptul că transparența nu poate împiedica discriminarea și nu poate asigura echitatea; aceasta nu este un panaceu pentru problematica evaluării. În mod ideal, ar trebui să se asigure posibilitatea de a opta pentru renunțarea la mecanismul de evaluare atunci când este posibil, fără prejudicii - în caz contrar, trebuie să se asigure mecanisme de contestare și rectificare a punctajelor. Acest lucru este deosebit de important în situațiile în care există o asimetrie a puterii între părți. Astfel de opțiuni de renunțare ar trebui să se asigure în etapa de proiectare a tehnologiei, dacă acest lucru este necesar pentru a asigura respectarea drepturilor fundamentale, fiind necesar într-o societate democratică.

d. Sistemele de arme autonome letale (LAWS)

(134) În prezent, un număr necunoscut de țări și industrii cercetează și dezvoltă sisteme de arme autonome letale, de la rachete capabile să țintească selectiv, la mașini care învață, cu abilități cognitive de a decide cu cine, când și unde să se lupte fără intervenție umană. Acest lucru generează preocupări etice fundamentale, cum ar fi faptul că ar putea să conducă la o cursă a înarmării incontrolabilă la un nivel fără precedent în istorie și să creeze contexte militare în care se renunță aproape în totalitate la controlul de către om, riscurile de funcționare defectuoasă nefiind abordate. Parlamentul European a solicitat elaborarea de urgență a unei poziții comune, cu caracter juridic obligatoriu, care să abordeze aspectele etice și juridice legate de controlul și supravegherea de către om, responsabilitatea și punerea în aplicare a legislației internaționale privind drepturile omului, a dreptului internațional umanitar și a strategiilor militare.[75] Amintind obiectivul Uniunii Europene de a promova pacea, consacrat la articolul 3 din Tratatul privind Uniunea Europeană, susținem și ne propunem să sprijinim Rezoluția Parlamentului din 12 septembrie 2018 și toate eforturile conexe privind sistemele de arme autonome letale.

e. Preocupări potențiale pe termen lung

---

73 Madary & Metzinger (2016). *Real Virtuality: A Code of Ethical Conduct. Recommendations for Good Scientific Practice and the Consumers of VR-Technology*. (Virtualitatea reală: un cod de conduită etică. Recomandări pentru buna practică științifică și pentru consumatorii de tehnologie a realității virtuale) Frontiers in Robotics and AI, 3(3).

74 Acest lucru se aplică, de asemenea, avatarurilor bazate pe IA.

75 Rezoluția Parlamentului European 2018/2752(RSP).

(135) Dezvoltarea IA este încă specifică domeniului și necesită cercetători științifici și ingineri umani bine pregătiți pentru a specifica cu exactitate obiectivele acesteia. Totuși, extrapolând în viitor, pe termen mai lung, pot fi emise ipoteze cu privire la anumite preocupări majore pe termen lung[76]. O abordare bazată pe riscuri sugerează că aceste preocupări ar trebui luate în considerare, având în vedere posibilele necunoscute încă necunoscute și evenimente „black swan".[77] Impactul ridicat al acestor preocupări, combinat cu incertitudinea actuală în ceea ce privește evoluțiile corespunzătoare, impune evaluări periodice ale acestor teme.

D. CONCLUZIE

(136) Prezentul document constituie Orientările în materie de etică pentru IA elaborate de Grupul de experți la nivel înalt privind inteligența artificială (AI HLEG).

(137) Recunoaștem impactul pozitiv pe care sistemele IA îl au deja și îl vor avea în continuare, atât din punct de vedere comercial, cât și social. Cu toate acestea, suntem totodată preocupați să asigurăm faptul că riscurile și alte efecte negative cu care sunt asociate aceste tehnologii sunt tratate în mod adecvat și proporțional ținând seama de aplicația IA. IA este o tehnologie atât transformativă, cât și perturbatoare, iar evoluția acesteia din ultimii ani a fost facilitată de disponibilitatea unor cantități enorme de date digitale, de progresele tehnologice majore în ceea ce privește puterea de calcul și capacitatea de stocare, precum și de inovația științifică și în materie de inginerie semnificativă în ceea ce privește metodele și instrumentele IA. Sistemele IA vor afecta în continuare societatea și cetățenii în moduri pe care nu ni le putem imagina încă.

(138) În acest context, este important să construim sisteme IA care să fie demne de încredere, deoarece oamenii vor putea să valorifice beneficiile acestora cu încredere și pe deplin numai dacă tehnologia, inclusiv procesele și oamenii din spatele tehnologiei, sunt de încredere. Prin urmare, la elaborarea prezentelor orientări, IA fiabilă a fost ambiția noastră fundamentală.

(139) IA fiabilă are trei componente: (1) IA ar trebui să fie legală, asigurând respectarea tuturor legilor și a reglementărilor aplicabile; (2) ar trebui să fie etică, asigurând respectarea principiilor și a valorilor etice și (3) ar trebui să fie solidă, atât din perspectivă tehnică, cât și socială, pentru a asigura faptul că, chiar dacă au intenții bune, sistemele IA nu provoacă daune neintenționate. Fiecare componentă este necesară dar nu este suficientă pentru a obține o IA fiabilă. În mod ideal, toate cele trei componente funcționează în armonie și se suprapun. În cazul în care apar tensiuni, ar trebui să depunem eforturi pentru a alinia aceste componente.

(140) În capitolul I, am prezentat drepturile fundamentale și o serie corespunzătoare de principii etice care sunt esențiale în contextul IA. În capitolul II, am enumerat șapte cerințe-cheie pe care sistemele IA ar trebui să le îndeplinească pentru a deveni o IA fiabilă. Am propus metode tehnice și metode fără caracter tehnic care pot ajuta la punerea lor în aplicare. În final, în capitolul III, am pus la dispoziție o Listă de evaluare pentru o IA fiabilă, care poate contribui la operaționalizarea celor șapte cerințe. În secțiunea finală, am oferit exemple de oportunități benefice și preocupări majore generate de sistemele IA, cu privire la care sperăm să stimulăm discuții suplimentare.

(141) Europa are o perspectivă unică, ținând cont de accentul pe care îl pune pe plasarea cetățeanului în centrul eforturilor sale. Acest accent este înscris în ADN-ul Uniunii Europene, prin tratatele pe care aceasta se întemeiază. Prezentul document face parte dintr-o viziune care promovează IA fiabilă; considerăm că această viziune ar trebui să reprezinte fundamentul pe care Europa își poate consolida poziția de lider în domeniul sistemelor IA inovatoare, de vârf. Această viziune ambițioasă va contribui la asigurarea prosperității umane a cetățenilor europeni, atât la nivel individual, cât și colectiv. Obiectivul nostru este să creăm o „inteligență

[76] Deși unii sunt de părere că inteligența artificială generală, conștiința artificială, agenții morali artificiali, superinteligența sau IA transformativă pot fi exemple de astfel de preocupări pe termen lung (care nu există în prezent), mulți alții consideră că aceste preocupări sunt nerealiste.

[77] Un eveniment „black swan" este un eveniment extrem de rar, cu un impact ridicat - atât de rar, încât ar putea să nu fie observat. Așadar, probabilitatea apariției poate fi estimată, în mod tipic, numai cu o mare incertitudine.

artificială (IA) fiabilă pentru Europa”, prin care beneficiile IA să poată fi valorificate de către toți cetățenii într-o manieră care să asigure respectarea valorilor noastre fundamentale: drepturile fundamentale, democrația și statul de drept.

GLOSAR

(142) Acest glosar se referă la orientări și este menit să ajute la înțelegerea termenilor utilizați în prezentul document.

Inteligență artificială sau sisteme IA

(143) „Sistemele de inteligență artificială (IA) sunt sisteme software (și, eventual, hardware) proiectate de oameni[78], care, dacă li se dă un obiectiv complex, acționează în dimensiunea fizică sau digitală, percepând mediul prin intermediul preluării datelor, prin interpretarea datelor structurate sau nestructurate colectate, prin raționarea cu privire la cunoștințe sau prin prelucrarea informațiilor obținute din aceste date și prin deciderea celei/celor mai bune acțiuni care trebuie întreprinse pentru a realiza obiectivul dat. Sistemele IA pot fie să utilizeze reguli simbolice, sau să învețe un model numeric și, de asemenea, își pot adapta comportamentul analizând modul în care mediul este afectat de acțiunile lor anterioare.

(144) Ca disciplină științifică, IA include mai multe abordări și tehnici, cum ar fi învățarea automată (exemple specifice de învățare automată fiind învățarea profundă și învățarea prin consolidare), raționarea automată (ce include planificarea, programarea, reprezentarea cunoștințelor și raționarea, căutarea și optimizarea) și robotica (ce include controlul, percepția, senzorii și actuatoarele, precum și integrarea oricăror altor tehnici în sistemele ciber-fizice).”

(145) Un document separat elaborat de Al HLEG, care explică definiția *sistemelor IA*, utilizată în scopul prezentului document, este publicat în paralel, cu titlul „O definiție a inteligenței artificiale (IA): principalele capacități și discipline științifice”.

Practicieni în domeniul IA

(146) Termenul „practicieni în domeniul IA” înseamnă toate persoanele fizice sau organizațiile care dezvoltă (inclusiv cercetează, proiectează sau furnizează date pentru IA), implementează (inclusiv pun în aplicare) sau utilizează sistemele IA, exclusiv cele care utilizează sistemele IA în calitate de utilizatori finali sau consumatori.

Ciclul de viață al sistemului IA

(147) Ciclul de viață al sistemului IA include etapa de dezvoltare (inclusiv de cercetare, proiectare, furnizare de date și testare limitată), implementare (inclusiv punere în aplicare) și utilizare a acestuia.

Posibilitatea auditării

(148) Posibilitatea auditării se referă la capacitatea unui sistem IA de a fi supus unei evaluări a algoritmilor sistemului, a datelor și a proceselor de proiectare. Aceasta constituie una dintre cele șapte cerințe pe care ar trebui să le îndeplinească o IA fiabilă. Acest lucru nu presupune neapărat că informațiile cu privire la modelele de afaceri și proprietatea intelectuală legate de sistemul IA trebuie să fie întotdeauna disponibile în mod liber. Asigurarea trasabilității și a mecanismelor de înregistrare din etapa inițială de proiectare a sistemului IA poate permite posibilitatea de auditare a sistemului.

Judecăți părtinitoare

(149) Judecățile părtinitoare reprezintă o înclinare a prejudiciilor către sau împotriva unei persoane, unui obiect sau unei poziții. Judecățile părtinitoare pot apărea în numeroase moduri în sistemele IA. De exemplu, în sistemele IA bazate pe date, cum sunt cele produse prin învățare automată, judecățile părtinitoare din colectarea datelor și formare pot conduce la un sistem IA care demonstrează judecăți părtinitoare. În IA bazată pe logică, cum ar fi sistemele bazate pe reguli, judecățile părtinitoare por apărea din cauza perspectivei pe care un inginer de cunoaștere ar putea să o aibă asupra normelor care se aplică într-un context specific. De asemenea, judecățile părtinitoare pot apărea din cauza învățării on-line și a adaptării prin interacțiune. Acestea pot apărea, de

[78] Oamenii proiectează sistemele IA în mod direct, dar pot utiliza și tehnici de IA pentru a optimiza proiectarea acestora.

asemenea, prin personalizare, și anume, în cazul în care utilizatorilor li se prezintă recomandări sau informații care sunt adaptate preferințelor utilizatorului. Acestea nu au legătură neapărat cu judecățile părtinitoare umane sau cu colectarea de date determinată de om. Judecățile părtinitoare pot apărea, de exemplu, prin contextele limitate în care un sistem este utilizat, caz în care nu există nicio posibilitate de a le generaliza la alte contexte. Judecățile părtinitoare pot fi bune sau rele, intenționate sau neintenționate. În anumite cazuri, judecățile părtinitoare pot conduce la rezultate discriminatorii și/sau inechitabile, indicate în prezentul document ca judecăți părtinitoare neechitabile.

Etică

(150) Etica este o disciplină academică, ce reprezintă un subdomeniu al filozofiei. La modul general, aceasta tratează aspecte precum „Ce este o acțiune bună?", „Ce valoare are viața omului?", „Ce este justiția?" sau „Ce înseamnă o viață bună?". În etica academică, există patru domenii majore de cercetare: (i) metaetica, care se referă în principal la semnificația și referința propoziției normative și la modul în care pot fi determinate valorile lor de adevăr (dacă există); (ii) etica normativă, mijloacele practice de determinare a unei acțiuni morale, examinând standardele pentru acțiunile corecte și greșite și atribuind o valoare acțiunilor specifice; (iii) etica descriptivă, care vizează o investigare empirică a comportamentului moral și a convingerilor oamenilor; și (iv) etica aplicată, care se referă la ceea ce suntem obligați (sau ni se permite) să facem într-o situație specifică (care, adesea, este nouă din punct de vedere istoric) sau într-un anumit domeniu de posibilități (adesea fără precedent în istorie) de acțiune. Etica aplicată tratează situații din viața reală, în care trebuie să se ia decizii sub presiunea timpului și adesea cu o judecată limitată. Etica IA este văzută adesea ca un exemplu de etică aplicată și se axează pe problemele normative generate de dezvoltarea, implementarea, punerea în aplicare și utilizarea IA.

(151) În cadrul discuțiilor etice, se utilizează frecvent termenii „moral" și „etic". Termenul „moral" face trimitere la modelele concrete, reale de comportament, la obiceiuri și convenții care se regăsesc în culturi, grupuri specifice sau la persoane fizice specifice la un anumit moment. Termenul „etic" face trimitere la o evaluare a acestor acțiuni și comportamente concrete din perspectivă sistematică, academică.

IA etică

(152) În prezentul document, termenul „IA etică" este utilizat pentru a indica dezvoltarea, implementarea și utilizarea IA care asigură respectarea normelor etice, inclusiv a drepturilor fundamentale ca drepturi morale speciale, principii etice și valori principale conexe. Aceasta este al doilea element dintre cele trei elemente principale necesare pentru realizarea IA fiabile.

IA centrată pe om

(153) Abordarea centrată pe om a IA urmărește să asigure faptul că valorile umane sunt centrale pentru modul în care sistemele IA sunt dezvoltate, implementate, utilizate și monitorizate, prin asigurarea respectării drepturilor fundamentale, inclusiv a celor prevăzute în Tratatele Uniunii Europene și în Carta drepturilor fundamentale a Uniunii Europene, toate acestea fiind reunite prin trimitere la un fundament comun care își are originile în respectarea demnității umane, ființa umană beneficiind de un statut moral unic și inalienabil. Aceasta implică, de asemenea, luarea în considerare a mediului natural și a altor ființe care fac parte din ecosistemul uman, precum și o abordare durabilă care permite prosperitatea generațiilor viitoare.

Red Teaming

(154) Red Teaming este o practică prin care o „echipă roșie" sau un grup independent provoacă o organizație pentru a îmbunătăți eficacitatea acesteia, asumându-și un rol sau un punct de vedere adversar. Această practică este utilizată în special pentru a contribui la identificarea și abordarea vulnerabilităților potențiale în materie de securitate.

Reproductibilitate

(155) Reproductibilitatea descrie dacă un experiment IA prezintă același comportament atunci când este repetat în aceleași condiții.

IA solidă

(156) Soliditatea unui sistem IA cuprinde atât soliditatea sa tehnică (adecvată într-un context dat, cum ar fi domeniul de aplicare sau etapa din ciclul de viață), precum și soliditatea sa din perspectivă socială (care asigură faptul că sistemul IA ia în considerare în mod corespunzător contextul și mediul în care acesta funcționează). Aceasta este esențială pentru a asigura faptul că nu pot apărea daune neintenționate, chiar dacă sistemul IA are intenții bune. Soliditatea este al treilea element dintre cele trei elemente necesare pentru realizarea IA fiabile.

Părți interesate

(157) Termenul „părți interesate" înseamnă toate părțile care cercetează, dezvoltă, proiectează, implementează sau utilizează IA, precum și cele care sunt afectate (în mod direct sau indirect) de IA - inclusiv, dar fără a se limita la, societăți, organizații, cercetători, servicii publice, instituții, organizații ale societății civile, guverne, autorități de reglementare, parteneri sociali, persoane fizice, cetățeni, lucrători și consumatori.

Trasabilitatea

(158) Trasabilitatea unui sistem IA se referă la capacitatea de a ține evidența datelor, a proceselor de dezvoltare și de implementare ale sistemului, în mod tipic prin intermediul identificării înregistrate și documentate.

Încredere

(159) Preluăm din literatură următoarea definiție: „Încrederea este văzută ca: (1) o serie de convingeri specifice care se referă la bunăvoință, competență, integritate și predictibilitate (convingeri bazate pe încredere); (2) disponibilitatea unei părți de a depinde de o altă parte într-o situație riscantă (intenție bazată pe încredere); sau (3) combinația acestor elemente."[79] Deși, de obicei, „încrederea" nu este o proprietate atribuită mașinilor, prezentul document urmărește să sublinieze importanța posibilității de a avea încredere nu numai în faptul că sistemele IA respectă legislația, că sunt etice și solide, ci și că această încredere poate fi atribuită tuturor persoanelor și proceselor implicate în ciclu de viață al sistemului IA.

IA fiabilă

(160) IA fiabilă are trei componente: (1) IA ar trebui să fie legală, asigurând respectarea tuturor legilor și a reglementărilor aplicabile; (2) ar trebui să fie etică, demonstrând respect pentru principiile și valorile etice și asigurând respectarea acestora și (3) ar trebui să fie solidă, atât din perspectivă tehnică, cât și socială, deoarece, chiar dacă au intenții bune, sistemele IA pot provoca daune neintenționate. IA fiabilă se referă nu numai la fiabilitatea sistemului IA în sine, ci cuprinde și fiabilitatea tuturor proceselor și actorilor care fac parte din ciclul de viață al sistemului.

Persoane și grupuri vulnerabile

(161) Nu există o definiție juridică general acceptată sau stabilită la scară largă a persoanelor vulnerabile, din cauza eterogenității acestora. Ceea ce constituie o persoană sau un grup vulnerabil este adesea specific contextului. Evenimentele de viață temporare (precum copilăria sau boala), factorii de piață (precum asimetria informațiilor sau puterea de piață), factorii economici (precum sărăcia), factorii legați de identitatea unei persoane (precum genul, religia sau cultura) sau alți factor pot juca un rol în acest sens. Carta drepturilor fundamentale a UE cuprinde, la articolul 21 privind nediscriminarea, următoarele motive, care pot constitui un punct de referință, printre altele: și anume, sexul, rasa, culoarea, originea etnică sau socială, caracteristicile genetice, limba, religia sau convingerile, opiniile politice sau de orice altă natură, apartenența la o minoritate națională, averea, nașterea, un handicap, vârsta și orientarea sexuală. Alte articole de lege abordează drepturile grupurilor specifice, pe lângă cele enumerate mai sus. Orice astfel de listă este neexhaustivă și poate fi modificată în timp. Un grup vulnerabil este un grup de persoane care au în comun una sau mai multe caracteristici ale vulnerabilității.

[79] Siau, K., Wang, W. (2018), Building Trust in Artificial Intelligence, Machine Learning, and Robotics (*Consolidarea încrederii în inteligența artificială, învățarea automată și robotică*), *CUTTER BUSINESS TECHNOLOGY JOURNAL* (31), S. 47–53.

Prezentul document a fost elaborat de membrii Grupului de experți la nivel înalt privind IA

enumerați mai jos în ordine alfabetică

Pekka Ala-Pietilä, președintele AI HLEG
IA Finlanda, Huhtamaki, Sanoma
Wilhelm Bauer
Fraunhofer
Urs Bergmann – coraportor
Zalando

Mária Bieliková
Universitatea Tehnică Slovacă din Bratislava
Cecilia Bonefeld-Dahl – coraportoare
DigitalEurope
Yann Bonnet
ANSSI
Loubna Bouarfa
OKRA
Stéphan Brunessaux
Airbus
Raja Chatila
IEEE Initiative Ethics of Intelligent/Autonomous Systems și Universitatea Sorbona
Mark Coeckelbergh
Universitatea din Viena
Virginia Dignum – coraportoare
Universitatea din Umeå
Luciano Floridi
Universitatea Oxford
Jean-Francois Gagné – coraportor
Element AI
Chiara Giovannini
ANEC
Joanna Goodey
Agenția pentru Drepturi Fundamentale
Sami Haddadin
Munich School of Robotics and MI
Gry Hasselbalch
Thinkdotank DataEthics și Universitatea din Copenhaga
Fredrik Heintz
Universitatea din Linköping
Fanny Hidvegi
Access Now
Eric Hilgendorf
Universitatea din Würzburg
Klaus Höckner
Hilfsgemeinschaft der Blinden und Sehschwachen
Mari-Noëlle Jégo-Laveissière
Orange
Leo Kärkkäinen
Nokia Bell Labs
Sabine Theresia Köszegi
Universitatea Tehnică din Viena
Robert Kroplewski
Avocat și consilier pentru guvernul polonez
Elisabeth Ling
RELX

Pierre Lucas
Orgalim – Europe's technology industries
Ieva Martinkenaite
Telenor
Thomas Metzinger – coraportor
Universitatea Johannes Gutenberg din Mainz și Asociația Universităților Europene
Catelijne Muller
ALLAI Țările de Jos și CESE
Markus Noga
SAP
Barry O'Sullivan, vicepreședinte al AI HLEG
University College Cork
Ursula Pachl
BEUC
Nicolas Petit – coraportor
Universitatea din Liège
Christoph Peylo
Bosch

Iris Plöger
BDI
Stefano Quintarelli
Garden Ventures
Andrea Renda
Facultatea din cadrul Colegiul Europei și CEPS
Francesca Rossi
IBM
Cristina San José
Federația Bancară Europeană
George Sharkov
Digital SME Alliance
Philipp Slusallek
German Research Centre for AI (DFKI)
Françoise Soulié Fogelman
Consultant IA
Saskia Steinacker – coraportoare
Bayer
Jaan Tallinn
Ambient Sound Investment
Thierry Tingaud
STMicroelectronics
Jakob Uszkoreit
Google
Aimee Van Wynsberghe – coraportoare
TU Delft
Thiébaut Weber
CES
Cecile Wendling
AXA
Karen Yeung – coraportoare
Universitatea din Birmingham

Urs Bergmann, Cecilia Bonefeld-Dahl, Virginia Dignum, Jean-François Gagné, Thomas Metzinger, Nicolas Petit, Saskia Steinacker, Aimee Van Wynsberghe și Karen Yeung au acționat în calitate de raportori pentru prezentul document.

Pekka Ala-Pietilä este președintele AI HLEG. Barry O'Sullivan este vicepreședinte, coordonând al doilea document al AI HLEG. Nozha Boujemaa, președinte până la 1 februarie 2019, care a coordonat primul document, a contribuit, de asemenea, la conținutul prezentului document.

Nathalie Smuha a oferit sprijin editorial.