unesco
Comisia Națională a
României pentru UNESCO

1956 - 2021
65
de ani - România în UNESCO

ETICA INTELIGENȚEI ARTIFICIALE

Cât de inteligent putem utiliza IA?

unesco
Comisia Națională a
României pentru UNESCO

COMISIA NAȚIONALĂ A ROMÂNIEI PENTRU UNESCO

Str. Anton Cehov nr. 8, sector 1, București, 011998
Tel.: +4 (021) 231.13.33/231.32.24
Fax: +4 (021) 230.76.36
cnr@cnr-unesco.ro
www.cnr-unesco.ro

**„Etica Inteligenței Artificiale.
Cât de inteligent putem utiliza IA?“**

**Această publicație face parte
din programul aniversar
„România și UNESCO, arc peste timp –
65 de ani CNR UNESCO“**

Mulțumiri Centrului Internațional de Excelență în Inteligența Artificială al Universității POLITEHNICA din București pentru contribuția deosebită.

Photo Credit:SHUTTERSTOCK

© Comisia Națională a României pentru UNESCO, 2021

Etica Inteligenței Artificiale. Cât de inteligent putem utiliza IA?

Inteligența Artificială va redefini noi teritorii în Știință

Madlen Șerban
Secretar General
Comisia Națională a României pentru UNESCO

În urmă cu trei ani, Audrey Azoulay, Directorul General UNESCO, a lansat un proiect ambițios: să ofere lumii un cadru de reglementare a utilizării Inteligenței Artificiale. Direcorul General a numit, în martie 2020, un grup ad-hoc compus din 24 de specialiști cu specializări multiple în domeniul Inteligenței Artificiale pentru a produce un text-cadru sub forma unei Recomandări UNESCO, care să țină cont de diferitele probleme contemporane globale.
Urmărind evoluția proiectului de Recomandare, Comisia Națională a României pentru UNESCO s-a aliniat la scopul și obiectivele pe care și le-a propus acesta, încercând să organizeze, la o scară mai mică, dezbaterile pe seama unui subiect atât de provocator precum Etica Inteligenței Artificiale.
Astfel, în 3 iunie 2021, CNR UNESCO a organizat colocviul online cu tema „Etica Inteligenței Artificiale.
Cât de inteligent putem folosi IA?", la care au participat numeroși membri IA comunității științifice de profil din România și un invitat străin, din partea Comisiei Naționale a Franței pentru UNESCO. În 7 octombrie, împreună cu Universitatea POLITEHNICA din București, un alt colocviu pe aceeași temă a fost organizat de catedra UNESCO în cadrul manifestării „Zilele UNESCO la POLITEHNICĂ".
Publicația de față, realizată împreună cu un grup de profesori și cercetători din domeniul eticii și IA, încheie seria de proiecte dedicate Eticii în Inteligența Artificială, fără a epuiza, însă, multiplele aspecte ale subiectului.
Această publicație este despre solidaritate intelectuală și morală! Este despre a fi acum și în viitorul care a început ieri. Este esența de a fi a UNESCO, care susține , alături de alte organizații internaționale, de state membre, asociate și alte state, dezvoltarea noilor tehnologii, prin cooperare și asigurarea transparenței. Inteligența artificială, ca orice tehnologie, oferă multiple avantaje.
Mulțumesc Florinei Pescaru, expert al CNR UNESCO pentru coordonarea întregului proiect!
Distinșii autori ai articolelor prezentate au evidențiat aplicabilitatea IA în diverse domenii. Dar, din păcate, tehnologia poate fi utilizată și în scopuri potrivnice umanității. Prin contribuția acestora, aflăm ce înseamnă IA explicabilă și de încredere. Care sunt provocările folosirii IA în medicină? – de exemplu, dacă algoritmii pentru interpretarea radiografiilor toracice sunt instruiți cu date de la pacienți în principal de sex masculin, rezultatele nu sunt la fel de precise atunci când sunt aplicate la interpretarea radiografiilor toracice ale pacientelor de sex feminin. Care sunt aplicațiile IA în industria aeronautică și în aeroporturi? Ce înseamnă etic și neetic în IA, cum pot fi folosite noile tehnologii pentru detectarea automată a postărilor sau mesajelor neetice, de bullying, misoginism, xenofobie? De asemenea, unele texte încearcă să răspundă la întrebări precum: Cum ne putem asigura că recomandările de etică sunt reflectate în practica implementării IA în diferite piețe? Care sunt efectele colaterale dăunătoare ale implementării IA în termeni de externalități morale, și nu de externalități economice? Unii autori au subliniat că legiferarea moralității nu este o sarcină simplă și rămâne una dintre problemele importante în filosofia morală, filosofia politică și filosofia dreptului. Suntem la început de drum. Știm că nu este doar un drum, ci o Cale a cunoașterii și a dezvoltării, pe care o vom parcurge împreună, în Lumea Nouă Interconectată.

Conținutul Recomandării UNESCO pentru Etica în Inteligența Artificială

Recomandarea își propune să pună în practică avantajele pe care Inteligența Artificială le aduce societății și să reducă riscurile pe care le implică folosirea IA. Recomandarea se asigură că transformările digitale promovează drepturile omului și contribuie la atingerea Obiectivelor de Dezvoltare Durabilă, abordând probleme legate de transparență, responsabilitate și confidențialitate, cu capitole de politică orientate spre acțiune privind guvernanța datelor, educație, cultură, muncă, sănătate și economie.

1. Protejarea datelor. Recomandarea solicită acțiuni dincolo de ceea ce întreprind firmele de tehnologie și guvernele, pentru a garanta persoanelor mai multă protecție prin asigurarea transparenței, mijlocirii și controlului asupra datelor lor personale. Aceasta prevede că toate persoanele ar trebui să poată accesa sau chiar șterge înregistrările datelor lor personale. Include, de asemenea, acțiuni de îmbunătățire a protecției și cunoașterii datelor și dreptul persoanei de a-și controla propriile date. De asemenea, susține capacitatea organismelor de reglementare din întreaga lume de a pune în aplicare acest lucru.

2. Interzicerea punctajului social și a supravegherii în masă. Recomandarea interzice în mod explicit utilizarea sistemelor IA pentru scorul social și supravegherea în masă. Aceste tipuri de tehnologii care sunt utilizate la scară largă, sunt foarte invazive, încalcă drepturile omului și libertățile fundamentale. Recomandarea subliniază că atunci când elaborează cadre de reglementare, statele membre ar trebui să ia în considerare faptul că responsabilitatea finală trebuie să revină întotdeauna oamenilor și că tehnologiile IA nu ar trebui să primească personalitate juridică în sine.

3. Ajutor în procesul de monitorizare și evaluare. De asemenea, Recomandarea pune bazele instrumentelor care vor ajuta la implementarea sa. Evaluarea impactului etic are scopul de a ajuta țările și companiile care dezvoltă și implementează sisteme IA pentru a evalua impactul acestor sisteme asupra indivizilor, asupra societății și asupra mediului. Metodologia de evaluare a gradului de pregătire ajută statele membre să evalueze cât de pregătite sunt în ceea ce privește infrastructura juridică și tehnică. Acest instrument va ajuta la îmbunătățirea capacității instituționale a țărilor și va recomanda măsuri adecvate pentru a se asigura că etica este implementată în practică. În plus, Recomandarea încurajează statele membre să ia în considerare adăugarea rolului unui ofițer independent de etică IA sau a unui alt mecanism pentru a supraveghea eforturile de audit și monitorizare continuă.

4. Protejarea mediului. Recomandarea subliniază faptul că actorii IA ar trebui să aprobe metodele IA eficiente din punct de vedere al datelor, al energiei și al resurselor, care să contribuie la asigurarea faptului că IA devine un instrument mai proeminent în lupta împotriva schimbărilor climatice și în abordarea problemelor de mediu. Recomandarea cere guvernelor să evalueze impactul direct și indirect asupra mediului pe parcursul ciclului de viață al sistemului IA. Aceasta include amprenta sa de carbon, consumul de energie și impactul asupra mediului al extracției de materii prime pentru sprijinirea producției de tehnologii IA. De asemenea, urmărește reducerea impactului asupra mediului al sistemelor IA și al infrastructurilor de date. Încurajează guvernele să investească în tehnologie ecologică și, dacă există un impact negativ disproporționat al sistemelor de inteligență artificială asupra mediului, Recomandarea indică faptul că acestea nu ar trebui utilizate.

Tehnologiile emergente, cum ar fi IA, și-au dovedit capacitatea imensă de a oferi servicii și bunuri. Cu toate acestea, efectele sale negative, care exacerbează o lume deja divizată și inegală, ar trebui controlate. Evoluțiile IA ar trebui să respecte statul de drept, evitând prejudiciile și asigurându-se că, atunci când acestea se produc, mecanismele de răspundere și de reparare sunt la îndemână pentru cei afectați.

Sursa: https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence

Inteligență Artificială pentru o viață mai bună

Florina Pescaru
Expert,
Subcomisia Știință,
Comisia Națională a României pentru UNESCO

Astăzi, inteligența artificială joacă un rol important în viețile a miliarde de oameni de pe pământ. Uneori neobservată, în laboratoare, pe internet, pe stații de cercetare orbitală, dar adesea cu consecințe profunde, ne transformă societățile și ne schimbă calitatea vieții.

Pandemia Covid-19, în timpul căreia învățământul online a devenit o necesitate, a accelerat imersia IA în existența noastră. Securizarea tranzacțiilor bancare în volume din ce în ce mai mari, în ultimii ani, a fost posibilă grație algoritmilor. Explozia de informații (big data) și folosirea accelerată a IA vor dinamiza toate sectoarele societății: vom avea servicii medicale mai bune, autovehicule și sisteme de transport mai sigure, servicii și produse mai ieftine și mai durabile. IA poate facilita accesul la educație și formare și ne poate îmbunătăți siguranța la locul de muncă prin preluarea de către roboți a sarcinilor considerate periculoase. IA ajută companiile pentru lucrul în regim remote și managementul de la distanță, îmbunătățind eficiența operațională. Cele mai multe organizații globale din domeniul sănătății se bazează pe software IA pentru sarcinile cotidiene.
Media relatează zilnic știri despre descoperiri majore în știință, precum aceea că folosind algoritmi de învățare automată, cercetătorii au identificat un nou antibiotic eficient sau că vor utiliza IA pentru a descoperi planete din afara sistemului nostru solar. Procesele sunt atât de disruptive încât apar temeri existențiale: unde se termină binele și unde începe răul produs de IA?
Henry Kissinger spunea în cartea recent apărută „The Age of IA And Our Human Future" că „inteligența artificială ar putea fi la fel de importantă ca și apariția armelor nucleare, dar și mai puțin previzibilă".
„Deciziile care afectează milioane de oameni ar trebui să fie corecte, transparente și contestabile. Aceste noi tehnologii trebuie să ne ajute să abordăm provocările majore din lumea noastră de astăzi, cum ar fi inegalitățile crescute și criza de mediu, și să nu conducă la adâncirea acestora", a declarat Gabriela Ramos, director general adjunct al UNESCO pentru Științe sociale și umane.

UNESCO consideră că reflecția globală asupra IA este de o importanță crucială pentru a se asigura că noile tehnologii, în special cele bazate pe IA, servesc binele umanității, contribuie la dezvoltarea durabilă și respectă drepturile omului și demnitatea umană. În acest spirit, la 24 noiembrie 2021, UNESCO a adoptat un instrument global de stabilire a standardelor pentru a oferi IA o bază etică solidă. Recomandarea privind Etica Inteligenței Artificiale nu numai că va proteja, ci și va promova drepturile omului și demnitatea umană și va fi o busolă de ghidare etică și o piatră de temelie normativă globală care să permită construirea unui respect puternic pentru statul de drept în lumea digitală.

Principii de Etică în Inteligența Artificială

Prof. dr. ing. Adina Magda Florea,
Universitatea POLITEHNICA din București

Inteligența Artificială (IA) are ca scop construirea de sisteme care manifestă un comportament rațional, analizează mediul și iau decizii autonome pentru realizarea unor sarcini specifice. Aplicațiile bazate pe IA sunt astăzi omniprezente, de la telefoane inteligente la roboți, mașini autonome și asistenți inteligenți, la traducere automată, sinteza opiniilor din volumele imense de texte și postări din rețelele sociale, pentru a da numai câteva exemple. Putem astfel spune că suntem deja în epoca Inteligenței Artificiale și aceasta este doar la început. Cu toate acestea, integrarea sistemelor de decizie bazate pe IA în viața curentă nu este simplă din numeroase motive. Există o serie de întrebări importante care trebuie puse despre viitorul nostru în care tehnologia și aplicațiile de IA vor fi omniprezente. Va deveni inteligența artificială mai „inteligentă" decât inteligența umană? Creează inteligența artificială bunăstare și oportunități pentru toți? Când trebuie să avem încredere în deciziile autonome ale IA și suntem în vreun fel puși în pericol de aceste decizii? În timp ce întrebarea legată de potențialul IA de a deveni sau nu mai inteligentă decât oamenii este încă deschisă pentru dezbateri ample, celelalte întrebări, și multe altele conexe, sunt acum printre preocupările centrale ale asociațiilor și organizațiilor internaționale și naționale, ale companiilor high-tech și ale societății civile, deopotrivă.

UNESCO a declarat că *„avem nevoie de o inteligență artificială centrată pe om, care trebuie să fie pusă în slujba oamenilor și nu invers"* și a propus dezvoltarea unui instrument global de stabilire a standardelor care să ofere IA o bază etică puternică, capabilă să protejeze și să promoveze drepturile omului și demnitatea umană. În iunie 2021, UNESCO a publicat un document despre *„Recomandări privind Etica Inteligenței Artificiale"*[1], care a fost înaintat spre adoptare Conferinței Generale la cea de a 41-a sesiune. Recomandările au coagulat 10 principii pentru realizarea IA etice:

- **Proporționalitate și cerința de a nu face rău** – utilizarea IA trebuie să fie proporțională pentru a atinge un anumit scop legitim și nu trebuie să încalce drepturile omului;
- **Siguranță și securitate** – IA trebuie să nu provoace daune umane sau materiale și să nu prezinte vulnerabilități la atacuri cibernetice;
- **Corectitudine și nediscriminare** – beneficiile IA trebuie să fie accesibile tuturor, iar aplicațiile de IA trebuie să evite discriminarea și părtinirea;
- **Sustenabilitate** – necesitatea de a evalua în permanență impactul uman, social, cultural, economic și de mediu al tehnologiilor de IA;
- **Dreptul la confidențialitate și protecția datelor** – dezvoltarea unor cadre adecvate pentru protecția datelor personale;
- **Supravegherea și determinarea umană** – un sistem de IA nu poate înlocui niciodată responsabilitatea umană finală;

[1] https://unesdoc.unesco.org/ark:/48223/pf0000377897

- **Transparență și explicabilitate** – capacitatea unui sistem de IA de a explica procesul de decizie;
- **Responsabilitate** – respectarea drepturilor omului și a libertăților fundamentale, protecția mediului și trasabilitatea sistemelor de IA;
- **Conștientizare și educare** – creșterea gradului de conștientizare și înțelegere a tehnologiilor de IA de către public, abordarea privind datele deschise și educația în IA accesibilă tuturor;
- **Guvernare adaptativă și colaborare** – participare deschisă, reglementări guvernamentale și internaționale atât pentru IA, cât și pentru utilizarea datelor.

Strategia privind inteligența artificială a Comisiei Europene vizează *„dezvoltarea și utilizarea IA etice și sigure, promovând o abordare centrată pe om în context global”*. Grupul de experți la nivel înalt pentru inteligența artificială (HLEG[2]) care sprijină implementarea Strategiei europene privind inteligența artificială a identificat patru principii etice care trebuie să stea la baza inteligenței artificiale de încredere: respectul pentru autonomia umană; prevenirea vătămării; corectitudinea; și explicabilitatea, și șapte cerințe cheie pentru realizarea IA de încredere: implicarea umană și supravegherea; robustețe tehnică și siguranță; confidențialitatea și guvernarea datelor; transparență; diversitatea, nediscriminarea și corectitudinea; bunăstarea mediului și a societății; responsabilitate.

Pentru a aplica aceste principii în viața reală, este nevoie de participarea tuturor actorilor implicați în întregul ciclu de viață al unui sistem de IA: cercetători și membri IA comunității academice, companii din sectorul privat, guverne, organizații interguvernamentale și societatea civilă. Inteligența artificială trebuie să sprijine oamenii și drepturile lor fundamentale și toți avem sarcina foarte dificilă de a dezvolta și implementa în continuare tehnologiile și sistemele de IA în moduri responsabile și robuste, de a construi reguli și reglementări care sunt acceptabile din punct de vedere politic, legal și etic, în beneficiul tuturor oamenilor și pentru realizarea unei societăți durabile și înfloritoare.

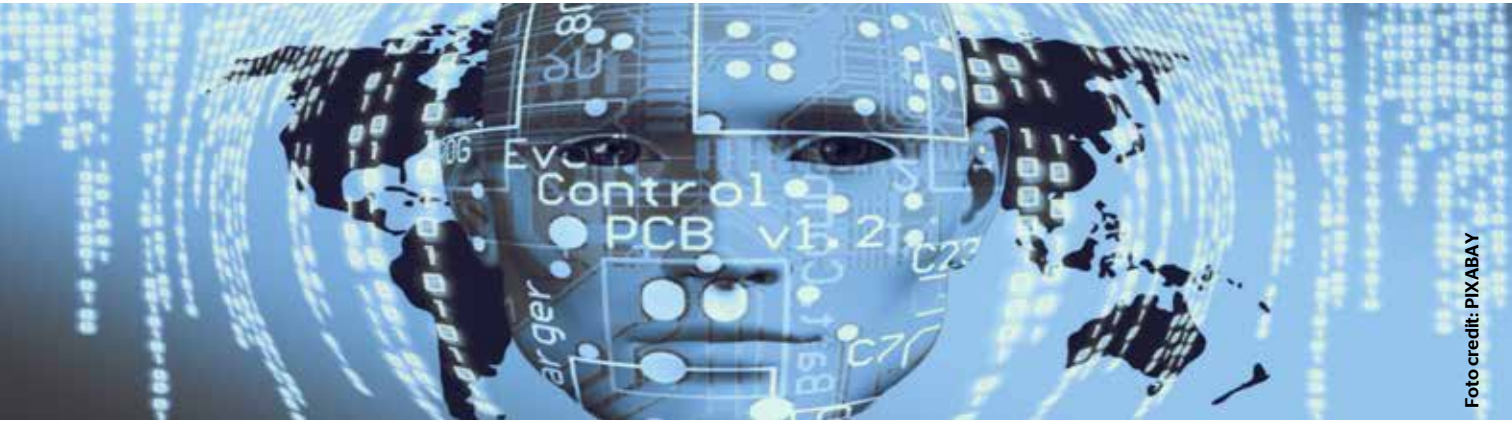


Foto credit: PIXABAY

[2] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Inteligența Artificială își marchează rolul în operațiunile aeroportuare și experiența pasagerilor

Prof. dr. ing. Sorin Eugen Zaharia,
Universitatea POLITEHNICA din București

Astăzi, transportul aerian reprezintă o parte indispensabilă a vieții moderne și este necesar pentru prosperitatea economică într-o economie globalizată. În 2019, în fiecare zi, s-au întâmplat următoarele: 12,5 milioane de pasageri au călătorit cu avionul; au fost efectuate 128.000 de zboruri regulate; au fost transportate mărfuri în valoare de 18 miliarde de dolari[1].

Aeroporturile lansează în mod constant noi inițiative digitale, iar inteligența artificială poate fi un atu extraordinar, permițând transportului aerian să optimizeze operațiunile și ajutând aviația să îmbunătățească continuu siguranța și securitatea, priorități pentru aeroporturi și pentru toți furnizorii de transport aerian. Aeroporturile lansează aplicații dedicate și acoperă domenii-cheie în care digitalizarea are un mare impact: operațiuni aeroportuare, securitate, luarea deciziilor în colaborare, soluții predictive și preventive, implicarea pasagerilor și comerțul cu amănuntul. În perioada de pandemie, soluțiile digitale au jucat și ele un rol important în câștigarea încrederii pasagerilor și în redresarea traficului.

Transformările digitale se referă în principal la automatizarea proceselor și implicarea pasagerilor, acțiuni care implică un management mobil al relațiilor cu clienții, tehnologia cloud, tehnologii block-chain, big data, Internet of Things (IoT) sau robotică. Un alt aspect important se referă la monitorizarea fluxului prin aplicarea de soluții predictive/preventive care se referă la geo-localizarea în interiorul aeroportului, managementul identității, managementul fluxului de pasageri sau identificarea prin radiofrecvență (RFID). În ceea ce privește eficiența operațiunilor aeroportuare, este importantă o abordare armonizată între diferiți actori. Aceasta implică sisteme cognitive bazate pe analiza datelor integrate de la sisteme de monitorizare de proces adecvate, cu scopul de a prezice și de a îmbunătăți procesele aeroportuare.

Influența culturii organizaționale în companiile aeronautice moderne poate reprezenta o cheie pentru succesul transformării digitale. Pe baza inteligenței artificiale (IA), designerii aeroporturilor folosesc mai multe modalități, prezentate mai jos, prin care încearcă să găsească cele mai bune soluții pentru îmbunătățirea operațiunilor și a experienței pasagerilor.

Robotul cu conducere autonomă experimentat pe aeroporturile din München, Frankfurt și Changi este un robot de transport și livrare bazat pe IA, care prin sistemul său de navigație integrat conduce pasagerii în tranzit la porțile de îmbarcare și îi poate ajuta să transporte bagaje de mici dimensiuni.

[1] ATAG, Aviation benefits beyond borders, 2021

Modalitățile în care este folosită IA sunt adesea invizibile pentru pasager. În culise, conducerile aeroporturilor intenționează să introducă modele de IA pentru a calcula prognozele privind numărul de pasageri. Aceste modele iau în considerare toate datele disponibile – de la vreme și trafic până la numărul de pasageri – și pot prezice câți pasageri vor fi într-o anumită locație la un moment dat. Calculând timpul necesar pasagerilor pentru a-și parcurge circuitul prin aeroport și înregistrându-și prezența în diferite puncte pe parcurs, aeroporturile pot fi mai bine pregătite pentru gestionarea vârfurilor de trafic și pot regla alocarea personalului în consecință sau pot redirecționa pasagerii către alte puncte de procesare.

Scoaterea aeronavelor din pozițiile de parcare, un alt proiect bazat pe IA, ar putea îmbunătăți **prognoza timpilor de escală pentru operațiunile la sol.** Metoda constă în a calcula, împreună cu partenerii externi, momentele de timp ale operațiunilor la sol ale sutelor de mii de zboruri și de a identifica factorii- cheie care ar putea afecta predicția operațiunilor de pe platformă. Analiza video folosită pentru a sprijini predicția poate transforma sistemele CCTV standard în sisteme inteligente și eficiente de detectare și alertă.

Oferind servicii de automatizare și folosind big data, Aeroportul Copenhaga și-a sporit eficiența, devenind astfel unul dintre aeroporturile din lume care utilizează tehnologii digitalizate pentru a reduce timpii de așteptare pentru recuperarea bagajelor sau pentru a accelera timpii de îmbarcare a pasagerilor[2]. Câștigătorul premiului pentru transformare digitală în 2017, Aeroportul Changi din Singapore, a reușit să creeze o identitate de aeroport unificată prin digitalizarea proceselor și a operațiunilor de bază și prin folosirea platformelor de date pentru rezolvarea problemelor și colaborarea între partenerii de afaceri. Din perspectiva managementului total al aeroportului, Changi a atins o mai bună anticipare operațională și de răspuns, și-a îmbunătățit planificarea resurselor, dobândind o identitate digitală unificată[3].

În ciuda procesului îndelungat de dezvoltare a unui model de IA puternic, beneficiile depășesc cu mult costurile. IA poate fi un mare atu, permițând aeroporturilor să scutească personalul existent de povara sarcinilor banale și repetitive, având mai mult timp pentru a-și asuma sarcinile specializate și mai importante[4]. IA aplicată de aeroporturi este fiabilă în aceeași măsură ca și datele pe care le utilizează, așa că prima provocare a aeroporturilor este să culeagă aceste date, asigurându-se că sunt suficiente și că sunt de înaltă calitate, în special pentru aeroporturile foarte mari care lucrează cu sute de parteneri. După depășirea acestui pas, implementarea IA trebuie să se asigure că datele sunt interpretate corect și că toate datele procesate sunt sigure, securizate și conforme cu GDPR.

Atunci când IA este implementată pe un aeroport, poate apărea o întrebare: personalul și pasagerii sunt circumspecți la AI? Fie că este vorba de o punctualitate sporită ca urmare a aplicării tehnologiei IA la activitățile din turnul de control, fapt ce contribuie la reducerea întârzierilor, sau de prognoze mai precise pentru pasageri, care vor ajuta la previziunea momentelor când culoarele de securitate vor fi cele mai aglomerate, tehnologiile sunt benefice în mod egal atât pentru pasageri, cât și pentru personalul aeroportului.

Utilizarea intenționată și etică a IA are potențialul de a transforma industria transportului aerian și de a rezolva o serie de provocări de afaceri într-un mod care nu a fost posibil până acum. Operarea mai efi-

[2] Kobenhavns Lufthavne, Europe's most efficient airport is in Copenhagen, 2016, https://www.cph.dk/en/ news/2016/6/europes-most-efficient-airport-is-in-copenhagen

cientă a aeroporturilor aduce beneficii pasagerilor, companiilor aeriene, personalului și comunităților locale. Ca și în cazul utilizării oricărei tehnologii, aeroporturile monitorizează progresele în aplicarea IA precum și reacția societății.

Aspectele prezentate sunt doar începutul aplicării IA în managementul aeroportuar, transportul aerian fiind gata să lucreze cu toate părțile interesate pentru a fructifica întregul potential al acesteia. Înscrierea aeroporturilor pe acest drum le situează deja într-o poziție foarte bună în aplicarea transformărilor digitale.

Foto credit: PIXABAY

[3] Jimenez, D. Z., Afuang, A., Rago, T., Tew,, K. L., Changi Airport Group wins Digital Transformer of the Year for Singapore at the 2017 IDC Digital Transformation Awards (DXa), Sep 2017, ttps://www.idc.com/getdoc.jsp?containerId=prAP43099417

[4] International Airport Review, Issue 01, February 2020

Etica Inteligenței Artificiale pentru Medicină

Prof. dr. Alexandru Scafa-Udriște,
Universitatea de Medicină și Farmacie „Carol Davila"

Prof. dr. ing. Adina Magda Florea,
Universitatea POLITEHNICA din București

Inteligența artificială (IA) transformă fiecare aspect al societății moderne și este probabil să aibă un impact enorm în următoarele decenii. Aplicațiile inteligenței artificiale pot contribui la promovarea creșterii economice, pot aduce mari beneficii societății și pot contribui la dezvoltarea individuală. În plus, IA demonstrează un potențial ridicat de contribuție la rezolvarea provocărilor globale, cum ar fi îmbunătățirea asistenței medicale sau lupta împotriva pandemiilor globale. IA poate îmbunătăți sănătatea personală și globală, poate reduce discrepanțele dintre sistemele de sănătate din diferite țări și poate contribui la asistență medicală personalizată. Cu toate acestea, utilizarea instrumentelor și aplicațiilor de IA în medicină ridică noi provocări din punct de vedere al implicațiilor etice ale implementării sale, de exemplu datele utilizate pentru antrenarea acestor aplicații sunt predispuse să determine noi tipuri de erori. Prin urmare, aplicațiile de IA pot crea riscuri, pot duce la vătămări neintenționate și pot crea provocări în legătură cu răspunderea și responsabilitatea juridică, subminând în același timp drepturile omului și procesele echitabile.

Mai multe aspecte ale utilizării etice a IA pentru sănătate au fost prezentate de diferite organizații. De exemplu, documentul *„Etică și guvernanță în inteligență artificială pentru sănătate"* [1], publicat în iunie 2021 de către Organizația Mondială a Sănătății, identifică principiile fundamentale care pot asigura funcționarea inteligenței artificiale pentru beneficiul tuturor țărilor și un set de recomandări privind utilizarea IA pentru sănătate.

Problemele etice din jurul IA în domeniul sănătății sunt complexe și se referă la diferite aspecte ale protecției drepturilor omului și asigurării bunăstării individuale. O viziune centrată pe om a IA pentru sănătate înseamnă că oamenii ar trebui să rămână în control deplin asupra sistemelor de îngrijire a sănătății și asupra deciziilor medicale. Mai mult, utilizarea datelor medicale pentru instruirea sistemelor de IA trebuie să protejeze aspectul privat și confidențialitatea datelor pacienților și să asigure consimțământul informat și

[1] https://www.who.int/publications/i/item/9789240029200

valid prin adoptarea cadrelor legale adecvate protecției datelor. Conform unui studiu recent, medicii dedică 62% din timpul lor pentru fiecare pacient în scopul analizei datelor clinice din dosarele de sănătate electronice (Electronic Health Record). Un sistem de IA poate ajuta medicii în acest proces, crescându-le eficiența și permițându-le să investească mai mult timp în îngrijirea pacienților. Volumele masive de date necesare construirii sistemelor de IA pot fi utilizate în scopuri pentru care datele nu au fost destinate inițial, datele pot fi reidentificate chiar dacă sunt anonimizate și au potențialul de a fi piratate și utilizate în scopuri dăunătoare sau comerciale.

Probleme de etică importante sunt generate de modul în care sunt luate deciziile de către un sistem de sănătate bazat pe IA, unde contează corectitudinea și relevanța datelor de intrare, iar în cazul în care acestea sunt incomplete, inexacte sau nereprezentative pentru diferite populații, corectitudinea algoritmilor aplicați pentru populația generală poate fi eronată. Părtinirea în compoziția datelor („data bias”) poate duce la perpetuarea sau chiar la amplificarea inegalităților

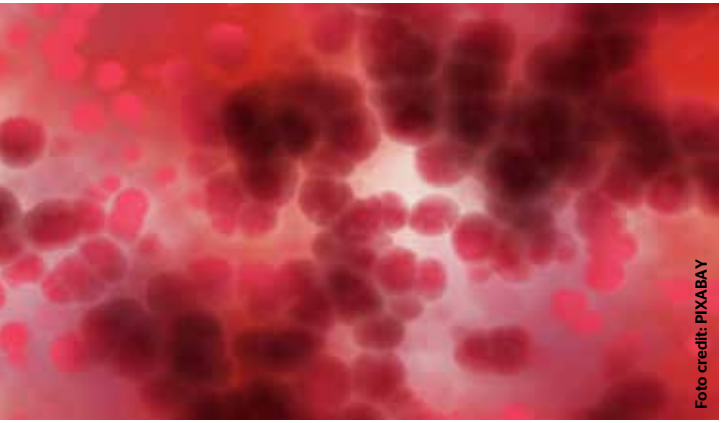


Foto credit: PIXABAY

bazate pe rasă, identitate de gen, vârstă sau caracteristici demografice și poate limita performanțele deciziilor de diagnostic sau tratament. De exemplu, dacă algoritmii pentru interpretarea radigrafiilor toracice sunt instruiți cu date de la pacienți în principal de sex masculin, rezultatele nu sunt la fel de precise atunci când sunt aplicate la interpretarea radiografiilor toracice ale pacientelor de sex feminin. Algoritmii de detectare a cancerului de piele instruiți în principal pe indivizi cu pielea deschisă la culoare au rezultate mai proaste la detectarea cancerului de piele care afectează pielea mai închisă. Un sistem de IA pentru recunoașterea activităților umane, atunci când este aplicat pentru a sprijini persoanele în vârstă în recuperarea unei boli, poate funcționa slab dacă este instruit doar pentru activitățile desfășurate de tineri. Responsabilitatea în luarea deciziilor este o problemă etică la fel de importantă: cine ar trebui să își asume responsabilitatea pentru erorile de diagnostic și tratament? Aceasta este o întrebare dificilă, deoarece multe sisteme de IA sunt opace, explicabilitatea și transparența fiind provocări actuale în cercetarea IA. Roboții de asistență pentru îngrijirea persoanelor în vârstă devin din ce în ce mai folosiți pe măsură ce roboții cu caracteristici umanoide avansează în performanță și scad ca preț. Totuși, dacă un astfel de robot dăunează unei persoane, cine va fi responsabil pentru acel prejudiciu: proiectantul ansamblului mecanic, inginerul de software sau medicul care a recomandat sarcinile asumate de robot? Utilizarea tehnologiilor de IA nu trebuie să aibă ca rezultat vătămări mentale sau fizice.

Toate aceste întrebări și provocări etice trebuie să fie discutate și să genereze răspunsuri corecte, deoarece IA pentru sănătate are un potențial imens, dar nu trebuie să submineze drepturile omului și accesul egal la îngrijirea sănătății.

Aspecte etice în Inteligența Artificială

Prof. dr. ing. Ștefan Trăușan-Matu,
Universitatea POLITEHNICA din București

Inteligența artificială (IA) este un domeniu puternic interdisciplinar, devenit omniprezent în viața tuturor, cu efecte majore asupra societății și a indivizilor. De aceea, asigurarea respectării principiilor eticii este un subiect foarte important. De exemplu, sunt multe aplicații cu IA care pot genera probleme de etică: agenții conversaționali, recunoașterea facială, extragerea de cunoștințe din conversații sau din postări pe e-mail sau rețele sociale, roboții de asistență a persoanelor în vârstă sau cu dizabilități, vehiculele autonome etc. În aceste aplicații se pot întâlni mai multe tipuri de probleme etice care pot afecta oamenii: generarea de replici neetice în conversații, părtinirea („bias"), luarea de decizii eronate, folosirea cunoștințelor extrase din postările de pe rețelele sociale în scopuri neetice. De exemplu, sunt cunoscute cazuri în care programe antrenate cu tehnici de IA au fost părtinitoare, de exemplu, în acordarea de credite la bancă sau eliberare condiționată din închisoare (conform *„Cartea albă privind Inteligența Artificială"* [1]). Profilurile utilizatorilor unei rețele sociale, construite cu tehnici de IA pot fi folosite în scopuri neetice. Pe de altă parte, inteligența artificială a unei mașini autonome, ce decizie să ia în caz de accident iminent care va afecta mai multe persoane: pe cine să accidenteze și pe cine nu?

Pentru a evita situațiile în care se încalcă etica în contextul IA, au fost elaborate documente ale Uniunii Europene, ale UNESCO, ale Consiliului Europei[2], articole ale mai multor cercetători și profesori de primă clasă din lume. De exemplu, în *„Cartea albă privind Inteligența Artificială"* o parte importantă este dedicată eticii. Totodată, Parlamentul European a publicat un document care evidențiază necesitatea unei abordări în IA centrată pe om[3]. Pe de altă parte, în cadrul unor mari companii, cum ar fi IBM sau Orange

[1] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[2] https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5
[3] https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf

Foto credit: PIXABAY

au fost elaborate documente de analiză a problematicii eticii în contextul IA.

Grupul de experți IA HLEG al Comisiei Europene a publicat o listă de criterii pentru evaluarea unei IA de încredere (*„Assessment List for Trustworthy Artificial Intelligence" - ALTAI*[4]). În aceste documente sunt evidențiate mai multe pericole care pot apărea în contextul exploziei aplicațiilor care folosesc inteligența artificială. Totodată au fost evidențiate și acțiuni care trebuie luate și identificați subiecții și entitățile care trebuie avute în vedere. Grupul de experți IA HLEG a identificat șapte cerințe esențiale pentru dezvoltarea de IA: implicare și supraveghere umană, robustețe tehnică și siguranță, respectarea vieții private și guvernanța datelor, transparență, responsabilitate, bunăstarea societății și a mediului, diversitate, nediscriminare și echitate.

Investigațiile asupra aspectelor etice ale IA ar trebui să răspundă la câteva întrebări: Care sunt implicațiile etice în utilizarea tehnologiei AI? Cum putem verifica îndeplinirea aspectelor etice? Care sunt posibilitățile de implementare ale roboților, agenților sau programelor de IA care iau în considerare principiile etice implicit sau explicit?

Detectarea unor situații de încălcare a eticii în contextul IA are două aspecte: evitarea generării de situații neetice datorate IA și folosirea acestor noi tehnologii pentru detectarea unor încălcări ale eticii de către alți agenți, umani sau artificiali, cum ar fi detectarea automată a postărilor sau mesajelor neetice, de exemplu, de bullying, misoginism, xenofobie etc.

Implementarea detectării situațiilor de încălcare a eticii este însă o problemă dificilă (dacă nu imposibil de rezolvat, în general), cu atât mai mult cu cât problematica eticii este dezbătută de milenii, și există mai multe teorii asupra ei – de exemplu, două din cele mai importante fiind teoriile deontologice (Kant, Kierkegaard și Nietzche) și teleologice (Aristotel și adepții utilitarismului). La primele, dimensiunea etică a unei acțiuni este dată de caracterul acțiunii respective, iar la cele teleologice doar de rezultatul acțiunii. Această distincție este făcută și în programele de IA: un model deontologic conține reguli, ontologii sau altă bază de cunoștințe care statuează ce este permis sau nu să se facă. În cazul abordării teleologice, se analizează doar etica rezultatului obținut, spre deosebire de primul caz nu sunt explicite criterii de etică folosite în acțiuni, situație des întâlnită în cazul IA sub-simbolică, bazată pe rețele neuronale.

[4] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

Bunul, Răul și Urâtul despre Sisteme IA Explicabile și de Încredere

Lector. dr. ing. Alexandru Sorici,
Universitatea POLITEHNICA din București

Există discuții recente despre inteligența artificială (IA) *explicabilă* și *de încredere* atât în reviste academice sau inițiative de cercetare finanțate de UE, cât și în media specializate. Noțiunea este prezentată ca o piatră de temelie pentru dezvoltarea și utilizarea etică a sistemelor de IA. Dar ce presupune cu adevărat IA explicabilă sau de încredere?

Grupul de experți la nivel înalt pentru IA al UE afirmă că ciclul de viață al unui sistem IA trebuie să se bazeze pe șapte principii. Printre acestea, *explicabilitatea* este împărțită în: *robustețe și siguranță tehnică, transparență și responsabilitate*.

Robustețea tehnică și siguranța înseamnă că sistemele IA trebuie să fie precise, fiabile și reproductibile.

Transparența înseamnă că sistemele IA, precum și echipele care le dezvoltă, trebuie să fie capabile să explice și să *comunice* elementele implicate în funcționarea lor: datele, sistemul în sine și modelul său de business.

Responsabilitatea este strâns legată de noțiunea de corectitudine. Necesită stabilirea unui mecanism de trasabilitate, atât înainte, cât și după dezvoltarea și punerea în funcțiune a unui sistem IA.

Bunul. Un raport recent din 2020 al Institutului de Cercetare Capgemini privind „IA și dilema etică" arată că există o conștientizare tot mai mare a necesității și practicării *explicabilității* în luarea automată a deciziilor, atât în rândul consumatorilor, cât și al directorilor de companii. S-au făcut eforturi pentru a explica funcționarea anumitor sisteme pe un limbaj pe care oamenii îl pot înțelege. Exemplele includ instituții financiare care creează „alternative digitale" ale unui profil de client, pentru a veni cu situații contrafactuale (tipurile de modificări ale variabilelor cheie care ar fi condus la un rezultat diferit) atunci când se uită la deciziile automate de respingere a unui împrumut. Google „Explainable AI" poate cuantifica modul în care fiecare punct de date contribuie la rezultat[1], în timp ce „InterpretML" de la Microsoft poate arăta factorii primari care dictează modul în care modelele sale de învățare automată iau o decizie[2].

În general, ponderea organizațiilor care își fac modelele de IA explicabile (cel puțin în sensul de a avea o narațiune interpretabilă) crește de la an la an.

Răul. Același raport Capgemini subliniază că, în timp ce *explicabilitatea* generală a sistemelor de IA înregistrează progrese, unele dintre elementele sale componente sunt în urmă. Consumatorii simt că devine mai greu *să știi* pe ce date a fost antrenat un sistem IA sau dacă sistemul ar putea oferi rezultate consecvente într-o manieră repetabilă.

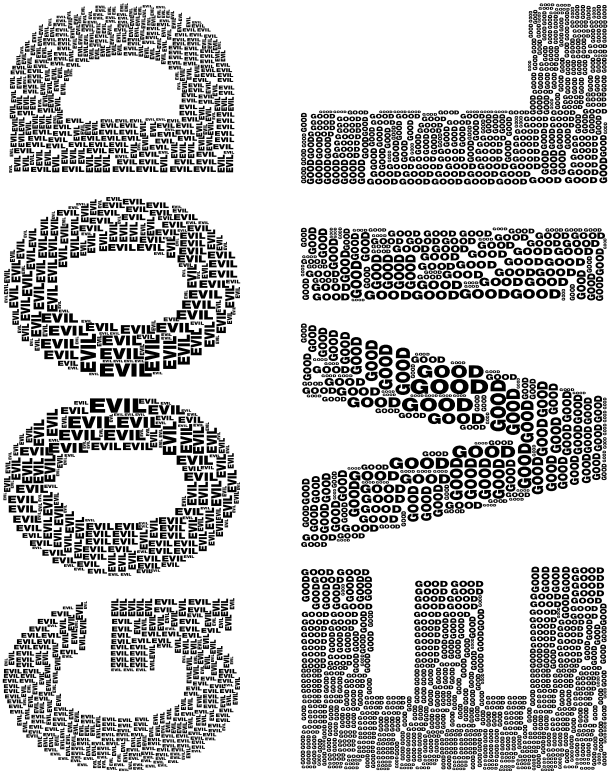
[1] https://www.researchworld.com/can-googles-new-explainable-ai-make-it-easier-to-understand-artificial-intelligence/
[2] https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability

Totodată, mai puțin de jumătate dintre organizațiile chestionate implementează un mecanism care auditează sistemele IA din perspectivă etică. Reproductibilitatea de la un capăt la altul (obținerea aceloraşi rezultate pentru intrări identice sau similare) nu este verificată. Totodată, organizațiile nu subliniază clar *cum* a fost construit sistemul IA, pe ce date a fost instruit și pe ce date a fost testat înainte de a fi pus în uz.

Urâtul. Problemele subliniate anterior pot fi parțial abordate prin eforturi de îmbunătățire a politicilor în jurul comunicării proceselor de proiectare, dezvoltare, testare și operare a sistemelor de IA. Cu toate acestea, o parte importantă a acestor probleme provine din *complexitatea tot mai mare* a modelelor de IA.

În timp ce pentru unele modele de învățare automată utilizate astăzi (de exemplu, arbori de decizie, modele automate de extracție a regulilor) se pot dezvolta instrumente tehnologice care permit transparență în funcționarea lor internă, același lucru nu poate fi afirmat pentru rețelele neurale complexe, cum ar fi modelul GPT-3 de procesare a limbajului natural, care are miliarde de parametri care pot influența rezultatul final. Într-un astfel de context, transparența *regulilor interne* ale modelului *nu* poate garanta *că regulile sunt interpretabile*, că au vreun înțeles pentru un om.

Concluzia. Problemele prezentate anterior arată de ce *explicabilitatea* sistemelor de IA este o chestiune necesară, dar în același timp complexă. Creșterea complexității modelelor de IA este motivul pentru care mulți cercetători se concentrează pe construirea de *instrumente tehnologice* pentru *interpretabilitatea* modelelor de IA, pentru a se asigura că oamenii pot *înțelege impactul* deciziilor luate de o soluție de IA. În timp ce dezvoltarea unor astfel de instrumente tehnologice este un proces în curs de desfășurare, organizațiile pot totuși spori robustețea, transparența și responsabilitatea sistemelor lor de IA prin măsuri la nivel de politică. Aceasta include aspecte precum: sublinierea clară a scopului preconizat al sistemului de inteligență artificială, asigurarea supravegherii umane, precum și punerea consumatorilor la conducerea interacțiunilor lor cu sistemul de IA, dându-le un control explicit asupra opțiunilor de confidențialitate.



Metavers, inteligență artificială și datele non-verbale: interferențe etice

Asist. univ. dr. Mihail-Valentin Cernea,
Academia de Studii Economice București

În 1992, Neal Stephenson publica romanul de literatură științifico-fantastică „Snow Crash", în care creionează viziunea unui nou Internet, în care prezența digitală, mediată de ecrane, tastaturi sau mouse-uri, este înlocuită de prezența fizică și mentală, un Internet care funcționează ca o metaforă a lumii reale – un metavers. Aproape 30 de ani mai târziu, Mark Zuckerberg anunță reorientarea eforturilor și resurselor Facebook către crearea unui astfel de metavers și schimbarea numelui companiei în „Meta", aducând problematica imaginată de Stephenson în realitate. Este greu să ne imaginăm că un astfel de spațiu virtual, dacă are succes și ajunge să fie populat de miliarde de utilizatori, poate să funcționeze în absența utilizării la scară largă a unor procese care să implice aplicații variate ale Inteligenței Artificiale în toate formele sale. Voi explora pe scurt, în cele ce urmează, potențiale probleme de ordin moral ce pot apărea la intersecția dintre metavers și IA, adică momentul în care algoritmii Big Tech vor avea acces nestrămutat la aproape toate aspectele ce compun o personalitate umană. Voi încheia intervenția introducând în ecuație complicațiile date de emergența așa-numitelor interfețe creier-calculator.

Înainte de toate, câteva detalii despre ce este, mai precis, un astfel de metavers: în principiu, vorbim de o realitate virtuală vastă, accesibilă prin caschete specializate, care ar trebui să funcționeze ca o extindere și ca un eventual înlocuitor al Internetului. În loc de site-uri și aplicații web, am călători cu propriul corp în spații care să îndeplinească cam aceleași funcții. De exemplu, o vizită pe site-ul unui mare retailer se transformă într-o călătorie într-un mall virtual în care putem interacționa cu o simulare 3D a produsului pe care am dori să-l achiziționăm. O întâlnire pe Zoom se transformă într-o șuetă între avataruri digitale într-un parc virtual. Visul lui Mark Zuckerberg nu este atât de îndepărtat pe cum ar putea părea, având în vedere că tipul de interacțiuni în VR pe care le-am descris mai sus sunt deja posibile cu tehnologia disponibilă pe piața de consum. Chiar Meta deține cea mai populară platformă de realitate virtuală de pe piață, Oculus Quest 2, cu aproximativ 4.6 milioane de unități vândute doar până în primul trimestru al anului 2021, conform Counterpoint Research. Între aplicațiile deja disponibile pe platforma Quest, munca într-un birou virtual comun este o realitate.

O bună parte din opinia publică a privit însă cu cinism anunțul celor de la Meta – mai degrabă o tactică de a schimba imaginea unei companii acuzată de o multitudine de probleme etice și legale. Cum utilizări variate ale inteligenței artificiale pe seturile de date strânse din activitatea utilizatorilor reprezintă,

practic, afacerea reală a fostei Facebook, merită să reflectăm asupra dilemelor etice specifice IA ce devin o realitate odată cu emergența metaversului.

Având în vedere că vorbim de o tehnologie în bună măsură deja existentă, modelată pe practicile actuale digitale, diferența dintre problemele morale specifice utilizării IA în industria IT, în general, și, respectiv, metavers în particular este doar una de grad, nu neapărat una de natură.

Ne putem pune problemele similare cu cele legate de rețele sociale sau jocuri video:

- respectarea dreptului la intimitate al utilizatorilor, în contextul unei prezențe virtuale mult mai accentuate;
- probleme legate de dependența pe care o poate da un astfel de spațiu, odată ce algoritmii care-i stau la bază determină preferințele unor consumatori vulnerabili;
- probleme legate de discriminarea algoritmică și modul în care nedreptățile din lumea reală vor fi replicate în lumea virtuală.

Odată ce realizăm dimensiunea și calitatea datelor personale care vor putea fi adunate de companii în metavers însă, urgența unei dezbateri morale despre protecțiile de care ar trebui să se bucure utilizatorii în viitoarele metaversuri ce vor răsări din toate colțurile Silicon Valley și efectul acestora asupra realității sociale devine o urgență.

Ce aduce în plus, în mod fundamental, pe piața de date, metaversul, este culegerea cu precădere a datelor non-verbale despre utilizatori: gesturile, mimica, mișcările trupului, unde își concentrează aceștia privirea, ba chiar și date mai profunde despre trăirile lor. Ultimele studii despre capacitatea de a identifica un utilizator pe baza acestui tip de date arată că unii algoritmi au nevoie de 5 minute de antrenament pentru a putea recunoaște 95% din indivizii despre care dețin date non-verbale (Miller et al. 2020). Astfel, în metavers, nu ne vom mai putea pierde în mulțime în fața IA.

Dacă adăugăm în această complexă ecuație tehnologică și dezvoltarea interfețelor creier-calculator care vor permite o comunicare neintermediată între mediile digitale și creierul uman, putem intui că utilizarea IA în metavers va crește masiv puterea giganților digitali de a cunoaște utilizatorii, dar și de a le determina preferințele.

Foto credit: PIXABAY

Etica utilizării Inteligenței Artificiale în aplicații militare

Dr. Cristian Ducu,
Centre for Advanced Research in Management and Applied Ethics

Probabil că cei mai mulți din cei care văd titlul *Etica utilizării inteligenței artificiale în aplicații militare* se vor gândi la roboți precum cei din filmele SF hollywoodiene. Suntem departe de momentul acela, dar există o mulțime de aplicații și scenarii de dezvoltare tehnologică care ne fac să ne punem din ce în ce mai multe întrebări de natură etică în această direcție.

Dezbaterea eticii IA nu se duce întotdeauna în sfera academică, unde argumentele par a fi cele mai valoroase, ci și în sfera politică, juridică și militară. Dilema majoră în această dezbatere pleacă de la argumentul folosit de unii politicieni și specialiști din zona militară, conform căruia nu ar trebui să existe (prea multe) restricții etice în inovarea și dezvoltarea aplicațiilor militare bazate pe IA (MinAI). Restricțiile de natură etică ar fi un dezavantaj major în fața altor țări care investesc în zona aceasta și nu țin seama de aceleași standarde morale, spun ei.

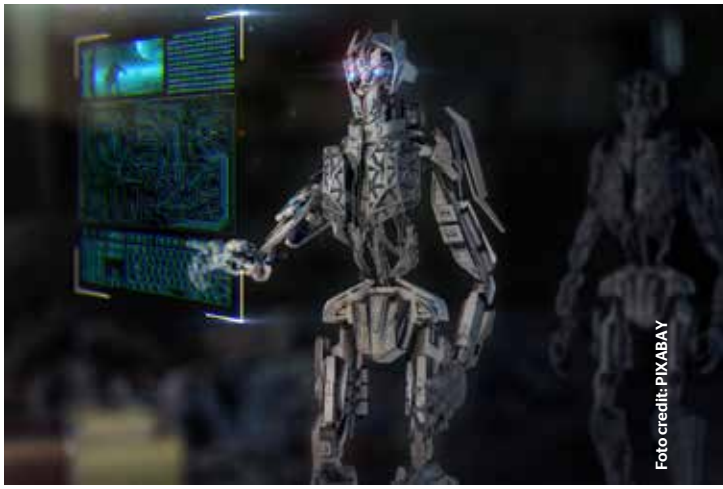


Foto credit: PIXABAY

Cel mai bun exemplu pentru o astfel de poziție a fost exprimată în 2021 de Nicolas Chaillan, primul Chief Software Officer al Departamentului de Apărare al Statelor Unite, care a acuzat, printre altele, că dezbaterile etice extensive în jurul Eticii IA țin pe loc Statele Unite din a investi, similar Chinei, în folosirea IA pentru a răspunde amenințărilor viitoare. Tot în 2021, Organizația Națiunilor Unite a cerut adoptarea unui moratoriu în ceea ce privește utilizarea IA în scopuri care pot aduce un prejudiciu drepturilor omului. Michelle Bachelet, Înaltul Comisar al ONU pentru Drepturile Omului, a menționat în mod special recunoașterea facială în timp real, din cauza problemelor legate de acuratețea recunoașterii, de discriminare și de protecție a dreptului la viață privată.

O astfel de tehnologie este folosită de serviciile de poliție din nenumărate țări și a fost utilizată de trupele americane în Afganistan, Irak și Siria pentru a identifica membrii grupărilor teroriste. Și alte țări, precum Israelul, China și Marea Britanie, folosesc

tehnologie similară în teatrele de operațiuni. În cazul tehnologiei de recunoaștere facială, care și-a demonstrat atât limitările, cât și beneficiile, pot fi argumentele etice un temei pentru a limita sau reconfigura utilizarea ei, pe viitor, în locuri critice precum frontierele, scenele unor atacuri teroriste sau zonele de conflict?

Lăsând deoparte chestiunea greutății argumentelor etice comparativ cu cele referitoare la securitate, există alte tehnologii militare care au la bază IA și care pun probleme etice și juridice serioase. De exemplu, există o amplă discuție la nivel internațional privind folosirea așa-numitelor „arme autonome" – echipamente militare care beneficiază de o minimă coordonare din partea unui operator uman. Imaginați-vă un vehicul aerian (UAV) comandat de algoritmi, care urmărește de la înălțime mare un autobuz școlar pe un drum de pământ dintr-o regiune controlată de teroriști. Respectivul UAV are capacitatea de a obține imagini de înaltă rezoluție și pe baza lor să identifice ținte, pe care ulterior să le angajeze. În acest caz, acel UAV, în lipsa unei conexiuni cu operatorul uman (din cauza faptului că lipsește acoperirea satelitară pentru comunicații), decide că în autobuz se află o țintă extrem de importantă – un lider terorist – și decide să atace cu o rachetă autobuzul. O primă întrebare de natură etică, în acest caz, ține de responsabilitate: cine poartă responsabilitatea, și, implicit, răspunderea, pentru o decizie „de viață și de moarte" luată de un algoritm? Multe persoane publice au făcut un apel, încă din 2012, ca așa-numiții „roboți ucigași" (*killer robots*) să fie interziși. În situația descrisă anterior, acea dezumanizare a deciziei cu privire la cine trăiește și cine moare a fost ușor combătută de oficiali militari care au admis, și chiar au insistat, asupra nevoii de control uman (*human governance of lethal decision-making*). Dar, adeseori, realitatea din teren este diferită.

O a doua întrebare de natură etică este legată de principiul proporționalității: este acea decizie de a ataca o țintă un curs al acțiunii dezirabil din punct de vedere al victimelor și distrugerilor estimate, comparativ cu amenințarea estimată? Dacă, de exemplu, în autobuzul școlar s-ar afla, pe lângă ținta urmărită, și cinci elevi, oare ar putea deveni decizia respectivă una moralmente justificabilă? Gândiți-vă la alt scenariu: un roi de 12 drone cu încărcătură activă (*weaponized drones*) au identificat mai multe ținte (teroriști) într-o clădire folosită drept școală. Teroriștii se întâlnesc într-o școală tocmai fiindcă ei mizează pe faptul că armata unui stat occidental nu va lua decizii care să conducă la victime în rândul copiilor. Doar că, de data aceasta, roiul de drone operează autonom și trebuie să decidă dacă se detonează în jurul clădirii respective. Ceea ce mai știm din comunicațiile interceptate este că acea întâlnire are ca scop pornirea unor atacuri coordonate împotriva unor obiective civile din mai multe țări occidentale. Cum ar trebui să cântărească din punct de vedere etic acel algoritm care coordonează activitatea roiului de drone?

Astfel de decizii etice sunt extrem de dificil de luat de IA în formele sale actuale, fără *input*-ul de la operatori umani. Foarte probabil că odată cu evoluția calculatoarelor cuantice, IA va depăși problema „dimensionalității" (capacitatea de a trata seturi de date comprehensive și de a lua decizii rapide pe baza informațiilor „învățate" sau „aflate") și, implicit, va putea să treacă la un nivel de decizie mult mai complex, care să conțină și o puternică latură morală. Totodată, acest uriaș pas făcut de tehnologie va conduce la noi provocări de natură etică, pe care la ora actuală încă nu le vedem.

Inteligență Artificială, externalități morale și reglementare permisivă

Dr. Mihaela Constantinescu,
Romanian Young Academy,
Centrul de Cercetare în Etică Aplicată (CCEA) - Facultatea de Filosofie, Universitatea din București

Pe măsură ce implementarea Inteligenței Artificiale (IA) bazate pe învățarea automată stratificată (deep learning) se confruntă cu tot mai multe provocări etice, necesitatea de dezvolta un cadru etic robust care să guverneze dezvoltarea viitoare a IA este din ce în ce mai presantă. Acest lucru a determinat organisme internaționale precum OECD, Comisia Europeană sau UNESCO să elaboreze documente prin care sunt propuse strategii, instrumente și mecanisme care integrează preocupările etice ale momentului.

În ciuda rezultatului pozitiv preconizat, mulți eticieni sunt încă reticenți față de modul în care vor fi utilizate în practică recomandările de etică, întrebându-se dacă toate acestea nu vor echivala pur și simplu cu un exercițiu de imagine (*ethics washing*) (Floridi, 2019). După cum subliniază unii cercetători, simpla existență a recomandărilor de etică nu va genera rezultatul așteptat în industrie, deoarece recomandările nu vor fi puse în practică în lipsa monitorizării adecvate (Hagendorf, 2020).

Așadar cum ne putem asigura că recomandările de etică sunt reflectate în practica implementării IA în diferite piețe? Care sunt cele mai bune metode pentru atenuarea efectelor dăunătoare ale utilizării IA? Sigur că nu există un răspuns simplu la astfel de întrebări. Totuși, în cele ce urmează voi schița un răspuns parțial printr-o propunere de înțelegere a impactului implementării IA prin prisma conceptului de externalități morale.

Foto credit: PIXABAY

Externalități morale ale implementării Inteligenței Artificiale

Pentru a discuta impactul implementării IA și relevanța recomandărilor de etică în practică, mă voi referi la conceptul de *externalități morale ale IA* și voi evidenția în special unele probleme legate de externalitățile morale negative ale IA. Pe scurt, prin externalități morale ale Inteligenței Artificiale înțeleg daunele și beneficiile morale colaterale implementării IA, suportate de o terță parte care nu are nicio responsabilitate pentru implementarea sau utilizarea IA. Această interpretare se bazează pe definiția folosită în domeniul eticii afacerilor, în care externalitățile morale negative se referă la „consecințe semnificative din punct de vedere moral care par să scape evaluării etice cu privire la ceea ce este datorat de către un actor – situații care depășesc capacitatea noastră de a atribui responsabilitate pentru prejudicii care pot fi prevenite" (Gowri, 2004: 40).

În mod evident, conceptul de externalități morale diferă de utilizarea sa inițială în economie - externalitățile morale nu sunt un subset, ci mai degrabă un fenomen paralel cu externalitățile economice deoarece, de exemplu, nu există un echivalent financiar pentru costurile sau beneficiile morale (Gowri, 2004). Drept urmare, abordarea economică a externalităților care urmărește internalizarea costului externalităților prin, de exemplu, taxe sau reglementări restrictive (*hard regulation*), nu are o strategie direct echivalentă atunci când vine vorba de externalități morale. Astfel, discutarea efectelor colaterale dăunătoare ale implementării IA în termeni de externalități morale, și nu de externalități economice, ne-ar putea ajuta să ne dăm seama mai bine de faptul că externalitățile IA nu pot fi pur și simplu echivalate cu unele costuri financiare și pur și simplu internalizate de către emitenți, de exemplu prin reglementări restrictive (*hard regulation*).

Cum poate fi de ajutor raportarea la conceptul de externalități morale pentru ca recomandările de etică privind dezvoltarea IA să fie într-adevăr respectate în practică, în diferite piețe?

Reglementarea implementării IA

În ciuda discuțiilor curente privind cele mai bune metode de reglementare a implementării IA (Reed, 2018; Taddeo și Floridi, 2018), nu există un consens asupra modului în care acest lucru poate fi realizat în practică. În primul rând, lucrurile sunt complicate de

însuși faptul că, adesea, nu este foarte clar ce anume trebuie reglementat (Almeida et al., 2020). În al doilea rând, legiferarea moralității nu este o sarcină simplă și rămâne una dintre problemele importante în filosofia morală, filosofia politică și filosofia dreptului (Hatzis, 2015). În al treilea rând, reglementarea este limitată în capacitatea sa de a capta proactiv multiplele consecințe dăunătoare generate de industriile cu evoluție rapidă, cum ar fi robotica și Inteligența Artificială.

Cu toate acestea, a lăsa deschisă industriei posibilitatea de auto-reglementare, în absența unui cadru etic comun și fără mecanisme de supraveghere, ar putea avea ca rezultat efecte dăunătoare imprevizibile asupra indivizilor, căci aceștia adesea nu au informațiile, instrumentele sau puterea pentru a exercita o presiune reală asupra pieței. Recomandările de etică adoptate și în curs de adoptare ar putea oferi într-adevăr acest cadru etic comun. Totuși, acestea riscă să fie inutile fără o operaționalizare adecvată în rezultate măsurabile, care să ofere un instrument de supraveghere pentru reglementatori și chiar pentru publicul larg.

IA, externalități morale și reglementare permisivă (*soft regulation*)

Externalitățile morale generate de implementarea Inteligenței Artificiale sunt o formă de externalități *soft* (folosesc termenul propus de Epstein (1997) atunci când vorbește despre externalități sociale) și, prin urmare, pot necesita mai degrabă un tip de reglementare permisivă (*soft regulation*), cum ar fi recomandările de etică. Însă, după cum am discutat mai sus, simpla existență a recomandărilor de etică nu este

PREZENTARE SCHEMATICĂ A EXTERNALITĂȚILOR MORALE ALE INTELIGENȚEI ARTIFICIALE

Nivel de dezvoltare	Aspecte etice	Externalități morale ale IA
Producția IA	*bias* la nivelul seturilor de date utilizate pentru antrenarea algoritmilor IA	polarizare, radicalizare
	colectarea datelor globale, extragerea și utilizarea *big data*	expunerea datelor private, reconfigurarea digitală a cunoașterii și adevărului
	utilizarea resurselor naturale și a piețelor de muncă cu costuri reduse pentru producția de hardware IA	exploatare umană, deteriorarea mediului
Natura IA	statutul moral al sistemelor autonome avansate de IA	scăderea controlului, autonomiei și responsabilității umane
Aplicații și utilizări ale IA	autoritate politică și control, libertate individuală	impact negativ asupra minorităților – părtinire, discriminare
	algoritmi de predicție (asigurări, sănătate, supraveghere etc.)	Acces restrâns/ restricționat la asigurări medicale, bias în recrutare, etc.
Impact pentru non-deținători și non-utilizatori de IA	utilizarea locală, regională, națională și globală a IA (în special în administrația publică, sănătate)	discriminarea non-utilizatorilor/non-deținătorilor, adâncirea inegalităților și decalajelor economice și sociale, acces inechitabil

suficientă pentru ca recomandările să fie adoptate și urmate în practică. O posibilă soluție este completarea recomandărilor cu instrumente suplimentare de guvernanță permisivă (*soft governance*), precum standarde etice (deschise) specifice (de exemplu, suita IEEE-P7000 și programele de certificare asociate; vezi IEEE, 2019). Având în vedere aplicarea lor practică (Winfield et al., 2021), astfel de standarde pot operaționaliza recomandările de etică privind IA și le pot face măsurabile. În plus, standardele au avantajul că pot fi aplicate prin reglementări restrictive în viitor, dacă situația o va impune (Theodorou & Dignum, 2020).

Pentru a atenua în mod corespunzător efectele dăunătoare ale implementării și utilizării IA, instrumentele combinate de reglementare permisivă oferite atât de recomandările de etică, precum și de standardele de etică, trebuie să ia în considerare o gamă largă de externalități morale ale IA. Tabelul alăturat oferă o schiță a externalităților morale de-a lungul următoarelor etape legate de implementarea IA: (1) producția IA (2) natura IA în sine (fie separat, fie parte a unui sistem IA robotizat) (3) aplicații și utilizări ale IA (4) impact pentru non-deținători și non-utilizatori de IA.

Pentru a concluziona, eforturile viitoare pentru reflectarea reglementării permisive în practica implementării IA vor trebui să ia în considerare modalități de a corela externalitățile morale ale IA cu rețeaua complexă și interconectată de responsabilitate morală asociată tuturor celor implicați de-a lungul întregului ciclu de implementare a Inteligenței Artificiale, cu o atenție deosebită acordată prejudiciilor morale colaterale ce se răsfrâng mai cu seamă asupra celor care se află *în afara* acestei rețele.

Bibliografie

Almeida, P., Santos, C.D., & Farias, J.S. (2020). Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 5257-5266.

Epstein, R. A. (1997). Externalities Everywhere: Morals and the Police Power. *Harvard Journal of Law and Public Policy*, 21, 61-69.

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology*, 32, 185–193.

Gowri, A. (2004). When Responsibility Can't Do It. *Journal of Business Ethics*, 54, 33-50.

Hagendorff, T. (2020). The Ethics of IA Ethics: An Evaluation of Guidelines. *Mind and Machines*, 30, 99-120.

Hatzis, A. N. (2015). Moral Externalities: An Economic Approach to the Legal Enforcement of Morality. In Aristides N. Hatzis & Nicholas Mercuro (Eds.), *Law and Economics: Philosophical Issues and Fundamental Questions* (pp. 226-244). London/New York: Routledge.

IEEE (2019). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. First Edition. Piscataway, NJ: IEEE Standards Association. Tech. rep.

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophy Transactions of the Royal Society*, 376.

Winfield AFT et al. (2021). IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and Artificial Intelligence*, doi: 10.3389/frobt.2021.665729.

Taddeo, R. & Floridi, L. (2018). How IA can be a force for good: An ethical framework will help to harness the potential of IA while keeping humans in control. *Science Review*, 361, 751-752.

Theodorou, A. & Dignum V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, 2, 10-12.

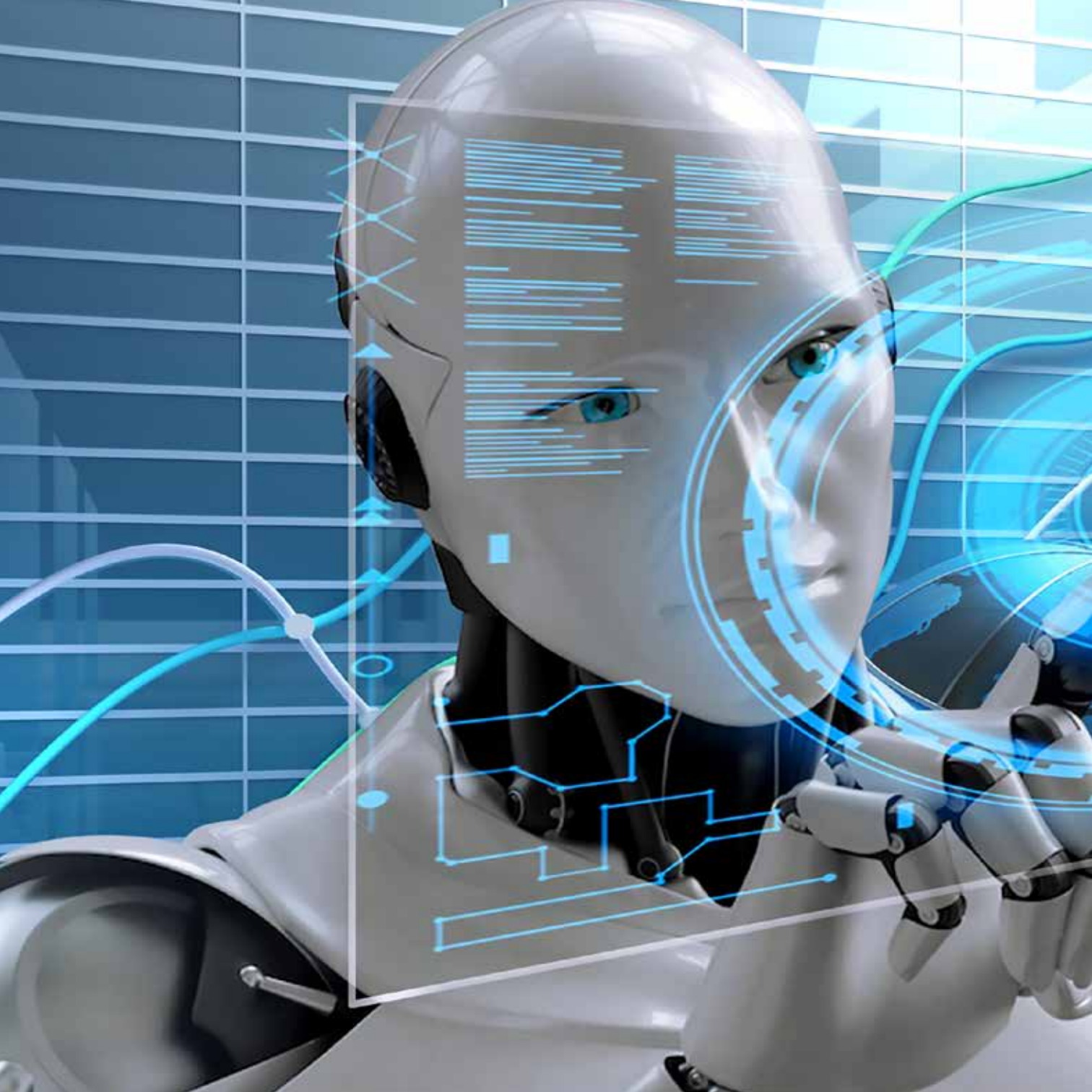


Foto credit: PIXABAY

AI