Sunt roboții buni sau răi?

Etica în IA: utilizări, limitări, și viitorul nostru comun

În articole precedente am vorbit pe larg despre sistemele de inteligență artificială (ce sunt, ce nu sunt și ce pot fi) și am identificat problemele majore ale acestor tehnologii și cum ne pot afecta drepturile fundamentale. Dacă doriți să explorați asta, găsiți articolul principal pe tema inteligenței artificiale și problemele ei sau cel despre discriminarea prin sisteme IA.

Introducere - tehnologie, etică și întrebări

Fiecare tehnologie poate fi folosită într-un anumit mod. Un mod bun sau un mod rău. Un topor trebuie folosit pentru a tăia lemne, nu oameni, iar aceeași regulă se poate aplica și sistemelor de inteligență artificială - respectând proporțiile comparației. Ele trebuie să vină în ajutorul oamenilor, nu împotriva lor. De aceea, etica în domeniul inteligenței artificiale este importantă – și un subiect recurent în orice discuție care vorbește despre soluțiile la problemele IA. **Este o discuție deschisă despre modul în care ar trebui să folosim inteligența artificială și spre ce scopuri ar trebui s-o îndreptăm încât să faciliteze înflorirea vieții omenești.**

Etica în domeniul inteligenței artificiale ar trebui să răspundă la întrebarea: care cale de acțiune este cea dreaptă, corectă, bună sau rea? Când vine vorba de inteligența artificială, lucrurile se complică, fiindcă inteligența artificială nu este un agent rațional și autonom, ca o ființă umană, așa că discuția despre etică în domeniul IA este o discuție despre consecințele sistemelor IA asupra vieților omenești. Cine este responsabil pentru un sistem IA care propagă discriminare de orice fel? Cum acționăm în conformitate cu o discriminare automatizată? Sunt roboții buni, sau răi?

În acest articol explorăm mai multe despre aceste întrebări, despre etica în domeniul inteligenței artificiale și, mult mai important, despre limitele eticii și, în final, despre cum etica nu poate înlocui reglementarea, ci doar o poate suplimenta. Discutăm și cum etica în domeniul IA poate facilita protecția drepturilor noastre fundamentale.

1. Etica aplicată în domeniul inteligenței artificiale

Etica normativă are o istorie lungă. Numele de etică vine din latinescul *ethos* care înseamnă un caracter bun, moral, drept. În general, etica normativă, adică cea care ține de *cum ar trebui* să ne comportăm (de aici normativitatea), se referă la întrebări legate de cum ar trebui să ducem o viață bună, ce e bun și ce e rău. Etica aplicată, despre care vorbim aici, diferă subtil. Etica aplicată înseamnă aplicarea unor principii morale despre bine și rău vieții de zi cu zi, sau unor situații specifice. De exemplu, patru principii morale fundamentale, din istoria filosofiei globale, des folosite în domeniul eticii aplicate sunt următoarele:

- Utilitarismul, adică maximizarea binelui celor mai mulți,
- Deontologismul, adică să acționăm în conformitate cu un percept moral universal - cum ar fi conceptul de ”datorie”- în orice context,
- Etica virtuților (după Aristotel), adică a acționa numai în conformitate cu acele acte morale care duc la înflorirea și desăvârșirea vieții omenești,
- Confucianismul, adică să nu faci altora ce nu ți-ai dori să ți se întâmple ție (regula de aur) sau să-i tratezi pe alții așa cum dorești să fii tratat la rândul tău

În contextul eticii aplicate în domeniul IA, etica se transformă în întrebări legate de care acțiune a unui sistem automatizat facilitează un rezultat pozitiv sau negativ, cine este responsabil pentru consecințele acțiunilor respective și ce înseamnă asta pentru noi, oamenii.

Etica aplicată în domeniul IA se referă, de obicei, la:

- seturi de reguli pentru conduita sistemelor IA,

- identificarea actorilor responsabili pentru acțiunile negative ale unui sistem IA (de obicei, aceștia sunt chiar producătorii sau furnizorii sistemelor IA),
- coduri de etică pentru comportamentul celor care le creează și implementează,
- declarații guvernamentale pe tema celui mai dezirabil comportament pentru sistemele IA,
- susținerea implementării și dezvoltării designului etic în domeniul tehnologiilor IA

Inteligența artificială este capabilă de decizii automatizate - câteodată și autonome. Asta înseamnă că aceste decizii pot fi puse sub semnul investigației de natură etică. Facilitează inteligența artificială o viață bună pentru oameni? Este folosită inteligența artificială într-un mod moral, pentru protecția oamenilor și pentru prosperitatea lor? Ar trebui anumite sisteme de inteligență artificială interzise, ca în cazul armamentului nuclear? De ce?

În cartea sa *AI Ethics* eticianul Mark Coeckelbergh sugerează că lucrurile nu sunt clare, când vine vorba de răspunsuri la întrebările de mai sus. Pe de o parte, inteligența artificială ne poate spori gradul de bunăstare prin eficientizarea unor industrii, precum cele medicale sau agroindustriale. Pe de altă parte, inteligența artificială, după cum am argumentat și noi într-un articol recent, poate propaga conținut discriminatoriu, sau chiar poate refuza persoane pe baza unor criterii obscure sau discriminatorii în sine. Iar interzicerea unor sisteme de IA, cum deja este inclusă în Regulamentul privind Inteligența Artificială (*AI ACT)* al Uniunii Europene, va fi percepută de mulți ca un atac la adresa inovației, chit că, de-a lungul istoriei, am stopat, cel puțin declarativ, dezvoltarea tehnologiilor rachetelor nucleare, de frica complet justificată a unui posibil război nuclear. Inteligența artificială nu este o tehnologie neutră, iar efectele ei pot pune în pericol persoane fizice și drepturile lor fundamentale.

Momentan, inteligența artificială tinde să:

- reducă indivizii la date/informații de antrenare diminuându-le astfel individualitatea,
- să sporească monitorizarea la adresa persoanelor fizice, atingându-le dreptul la viață privată,
- să perpetueze stereotipuri și prejudecăți din datele de antrenare nediversificate,
- și să exacerbeze legăturile dintre unele grupuri de oameni și acțiuni de un anumit fel, perpetuând astfel prejudecăți deja puternic înrădăcinate în societățile contemporane

Conform lui Coeckelbergh, o problemă fundamentală pentru etica în domeniul inteligenței artificiale este cea a corelațiilor false (*spurious corelations*). Corelațiile false se referă la faptul că, într-un set de date gigantic, pe baza cărora funcționează sistemele de inteligență artificială, vor exista corelații arbitrare între anumite date, fără vreo legătură de cauzalitate reală. De exemplu, dacă avem un set gigantic de date și o mașinărie statistică, cum sunt unele sisteme IA, putem trage legături, adică corelații, între X și Y fără ca X și Y să aibă vreo legătură reală.

Putem ajunge la concluzii, la modul absurd, cum ar fi că creșterea economică într-o societate este strâns legată de numărul de oameni care ascultă Pink Floyd, sau că nivelul de educație al unei țări este strâns legat de numărul de prieteni de pe Facebook pe care-l are fiecare individ. Evident, acestea sunt exemple fictive, menite a ilustra ideea de mai sus.

Doar că așa funcționează cele mai multe mașinării statistice din domeniul inteligenței artificiale (cum ar fi *machine learning* și *neural networks)* și pe baza seturilor largi de date *(big data*) se creează tot felul de corelații aberante, care duc la rezultate ciudate, posibil discriminatorii dacă mașinăria produce corelații între grupul X și infracțiunea Y, când nici un fel de corelație de genul nu există în realitate. Am vorbit despre asta într-un alt articol pe tema discriminării automatizate.

Așadar, realitatea conturată de inteligența artificială este una confuză din punct de vedere etic: eficientizare și inovație pe de o parte care ar trebui, în teorie, să ne îmbunătățească viețile, corelații false, discriminare, monitorizare și procesare de date într-un mod ilegal pe de altă parte. Cum ne poate ajuta etica aplicată în acest context?

1.1. Beneficiile eticii aplicate în domeniul IA

Etica aplicată ne poate ghida comportamentul și rațiunea înspre construirea unor sisteme de inteligență artificială care să faciliteze prosperitatea umană, longevitatea și protecția drepturilor fundamentale, prin coduri de etică, sau alte sisteme de reguli informale mai sus menționate. **În cartea sa *The Fourth Revolution*, eticianul și filosoful informației Luciano Floridi argumentează că etica este deseori văzută ca fiind ceva restrictiv la adresa comportamentului – o numește *etică negativă.* Etica negativă spune: nu face acțiunea X, pentru că Y. Asta este etica celor mai multe coduri de etică. O astfel de etică este**

limitativă, sugerează Floridi. Vede doar negativul din etică, și nu vede posibilitatea de îmbunătățire a conduitei unui sistem IA. În schimb, avem nevoie de o nouă etică – *o etică pozitivă.* Etica pozitivă pune accentul pe deliberare, calcularea și cântărirea acțiunilor, pe facilitarea raționamentului etic, pe gândirea morală, pe privilegierea vieții și prosperității umane. El gândește etica pozitivă ca trebuind să fie parte integrală a designului tehnologiilor noi, cum ar fi inteligența artificială.

Etica pozitivă, propusă de Floridi, ne poate opri din logica profitului-cu-orice-cost, starea de facto a multor entități publice și private, pentru reflecție asupra propriul comportament și asupra consecințelor comportamentului nostru. **Etica pozitivă a lui Floridi ne poate spune faptul că sistemele de inteligență artificială din momentul actual nu sunt construite pentru prosperitatea generală umană, ci pentru prosperitatea economica a unor entități publice sau private. Iar astfel de acțiuni sunt imorale, dacă definim scopul tehnologiei de tip IA ca fiind îmbunătățirea și ameliorarea vieții omului. Din punctul de vedere al eticii pozitive despre care vorbește Floridi, corelațiile false sunt imorale, fiindcă riscă să perpetueze stereotipuri prin asocieri false dintre populații și anumite tipuri de infracțiuni, printre multe altele.** Asta ar trebui să ne facă să considerăm stoparea sau schimbarea sistemelor de IA care folosesc corelații false și *big data*. Uniunea Europeană a înțeles importanța eticii pozitive despre care vorbește filosoful Luciano Floridi. În documentul faimos de la Comisia Europeană din 2019, numit ”Orientări în materie de etică pentru o inteligență artificială (IA) fiabilă” patru principii de *etică pozitivă* domină discuția despre construcția sistemelor de inteligență artificială:

1. respectarea autonomiei oamenilor
2. prevenirea daunelor,
3. echitatea, și
4. explicabilitatea.

Există o tendință ca fiecare nouă tehnologie să se transforme într-o armă. Dinamita, inventată de Alfred Nobel, trebuia să ajute minerii să spargă pereții minelor, dar oamenii i-au întrebuințat o altă utilizare, bine cunoscută actualmente. Sistemele de inteligență artificială nu sunt diferite de cazul dinamitei, iar cele enumerate de mai sus, adică corelațiile false, discriminarea automatizată, sporirea monitorizării automatizate, perpetuarea prejudecăților și altele arată că sistemele IA pot fi arme fără voia lor, care n-au fost proiectate pentru așa ceva

dar care sporesc suferința umană și afectează astfel drepturile fundamentale. **Așadar, ele trebuie regândite, din momentul conceperii, de jos în sus și de sus în jos, ca tehnologii menite spre a spori viața omului, spre a o ameliora, a o îmbunătății și a o face, în general, mai bună.**

2. Limitele eticii în domeniul inteligenței artificiale

Etica aplicată în domeniul inteligenței artificiale are niște limitări evidente. În primul rând, etica este văzută, deseori, ca o reglementare de mâna a doua, fără cadru legal. Ea este percepută ca fiind un set de recomandări care nu constrâng prin pedepse sau recompense, și, deseori, este văzut ca fiind complet opțională – ba chiar o barieră împotriva inovației. Pe lângă aceste limitări, evidente la prima vedere, există și o latură mai puțin cunoscută despre modul în cadre industria de inteligență artificială folosește etica aplicată.

Fenomenul "spălare de etică" (*ethics washing)* ca în cazul spălării banilor, redat în Tagespiegel de Thomas Metzinger, filosof care a luat parte la întocmirea documentului "Orientări în materie de etică pentru o inteligență artificială (IA) fiabilă" din 2019 este deja bine cunoscut de eticienii din domeniu. **Spălarea de etică se referă la faptul că oameni din industria inteligenței artificiale organizează discuții în jurul ideii de etică aplicată în domeniul IA pentru a câștiga timp și pentru a întârzia procesul de reglementare.**

Conform lui Thomas Metzinger, industria din domeniul IA acționează mai rapid și mai eficient decât sectorul academic și există riscul ca, la fel ca în cazul știrilor false "să avem acum și o problemă cu etica falsă, inclusiv o mulțime de ecrane de fum și oglinzi conceptuale, filosofi industriali bine plătiți, sigilii de calitate inventate de ei înșiși și certificate nevalidate pentru "IA etică fabricată în Europa"". Astfel, discuția despre etica aplicată în domeniul IA poate fi folosită pentru a distrage atenția spre beneficiul industriei de inteligență artificială prin apelul la simbolistica eticii (filosofi, certificate, discuție, etc) dar fără substanța ei.

Luciano Floridi, filosoful mai devreme menționat, a întocmit o taxonomie a modurilor prin care etica poate fi folosită în mod înșelător de industrie, pe care o vom reda aici tradusă din engleză și adaptată pentru a fi mai ușor de înțeles:

- Cumpăratul digital de etică (digital ethics shopping) definit ca: greșeala de a alege, adapta sau revizui ("mixing and matching") principii etice, orientări, coduri, cadre sau alte standarde similare (în special, dar nu numai, în etica inteligenței artificiale), dintr-o varietate de oferte disponibile, pentru a adapta anumite comportamente (alegeri, procese, strategii etc.) preexistente și, prin urmare, pentru a le justifica ulterior în loc să implementeze sau să îmbunătățească noile comportamente prin compararea lor cu standardele etice publice. *De exemplu, unele companii susțin că lucrează cu echipe care se ocupă cu definirea sau evaluarea eticii din spatele algoritmilor de AI folosiți, însă cetățenii nu au cum să știe dacă principiile etice folosite de companii diferite sunt compatibile, similare sau dacă sunt în linie cu etica lor, personală.*

- Bluewashing cu etică (ethics bluewashing) definit ca: practica greșită de a face afirmații nefondate sau înșelătoare cu privire la valorile și beneficiile etice ale proceselor, produselor, serviciilor sau altor soluții digitale sau de a implementa măsuri superficiale în favoarea acestora, pentru a părea mai etic din punct de vedere digital decât este cazul în realitate. *De exemplu, atât OpenAI (care deține ChatGPT) cât și Anthropic (care deține Claude) au anunțat că vor redirecționa întrebările pe teme electorale către "autorități pe subiect" (site-uri precum CanIVote.org sau TurboVote.org), dar nu au spus când vor implementa măsura asta, iar publicația Proof a descoperit că nu s-au ținut de cuvânt, încă. Aceste măsuri sunt menite să scadă riscul de dezinformare în timpul campaniilor electorale.*

- Lobby în domeniul eticii digitale (digital ethics lobbying), definit ca: greșeala de a exploata etica digitală pentru a întârzia, revizui, înlocui sau evita o legislație bună și necesară (sau aplicarea acesteia) cu privire la proiectarea, dezvoltarea și implementarea de procese, produse, servicii sau alte soluții digitale. *De exemplu, companiile de tehnologie care produc soluții de tipul inteligență artificială și-au crescut bugetele pentru lobby și și-au îndesit întâlnirile cu europarlamentari înaintea votului pe subiectul Regulamentului Privind Inteligența Artificială (AI Act), un proiect legislativ al Uniunii Europene menit să reglementeze aplicațiile de IA.*

- Dumping (concurență neloială) etic digital (Digital ethics dumping), definit ca: practica greșită care constă în exportul de activități de cercetare privind procesele,

produsele, serviciile sau alte soluții digitale, în alte contexte sau locuri (de exemplu, de către organizații europene din afara UE), în moduri care ar fi inacceptabile din punct de vedere etic în contextul sau locul de origine și importul rezultatelor unor astfel de activități de cercetare neetice. *Un exemplu este legat de mașinile care au implementat un algoritm pilot care poate conduce mașina fără intervenția șoferului. În 1968 a fost semnată Convenția de la Viena cu privire la trafic, însă Statele Unite ale Americii nu au semnat actul. Acest act impune ca șoferii să aibă control și responsabilitate depline în timpul șofatului. Drept rezultat, majoritatea mașinilor care folosesc pilot automat sunt testate pe drumurile din Statele Unite ale Americii, unde au produs deja victime.*

- Evitarea eticii (ethics shirking) definit ca: greșeala de a face din ce în ce mai puțină "muncă etică" (cum ar fi îndeplinirea îndatoririlor, respectarea drepturilor și onorarea angajamentelor) într-un anumit context, cu atât mai puțin cu cât randamentul unei astfel de munci etice în acel context este perceput în mod eronat ca fiind mai mic. *Cartea "Ethics Dumping" (disponibilă în întregime online) detaliază câteva exemple ale unor astfel de practici, mai ales legate de domeniul medical. Multe dintre studiile de caz urmăresc colaborări între nordul și sudul global, care sunt marcate de diferențe extreme în ceea ce privește veniturile și puterea disponibile, precum și un trecut colonial.*

Aceste șase mijloace de abuz ale eticii aplicate în domeniul inteligenței artificiale sunt, în general, modurile prin care entitățile care produc sau folosesc inteligența artificială se eschivează de reglementarea reală a industriei lor și mențin o iluzie a comportamentului etic și prosocial. Desigur, lista nu este completă, este, momentan, orientativă dar utilă pentru a înțelege atât limitele eticii în domeniul inteligenței artificiale dar și modul în care discuția despre etică, în general, poate fi folosită pentru a distrage atenția, pentru campanii de marketing sau PR sau pur și simplu pentru evitarea reglementării.

3. Reglementarea nu este egală cu etică. Reglementarea bate etica.

Etica nu constrânge persoanele fizice și juridice spre un comportament pozitiv, care facilitează sporirea și înflorirea vieții omenești, deoarece codurile de etică sunt doar la nivelul recomandărilor, fără vreun cadru legal în spatele lor. În schimb reglementarea nu este

opțională, ci obligatorie, și impune constrângeri reale, prin lege, care facilitează un anumit tip de comportament în detrimentul altuia.

Etica aplicată în domeniul IA este opțională, ca orice fel de etică. Etica, prin definiție, nu poate fi obligatorie. Nimeni nu are monopol asupra a ce este bun sau rău. În absența unor constrângeri reale, a unor pedepse sau recompense pentru acțiunile etice ale celor din industrie și ale sistemelor de inteligență artificială, etica este doar un set de percepte înșirate pe o foaie, fără vreo putere reală. Așadar, reglementare este mai puternică și mai dezirabilă în combaterea acțiunilor imorale sau periculoase ale furnizorilor și utilizatorilor de sisteme de inteligență IA.

De exemplu Regulamentul UE privind Inteligența Artificială (*AI ACT, încă nefinalizat la momentul publicării acestui articol)* impune anumite obligații pentru furnizorii de sisteme de inteligență artificială pe care un simplu cod de etică – sau o discuție despre virtuțile lui Aristotel – nu le poate impune, ci doar recomanda. Iar recomandările au această tendință de a fi ignorate, mai ales când este vorba despre o industrie gigantică de ordinul sutelor de miliarde cum este cea a inteligenței artificiale.

Prin urmare, reglementarea bate etica în domeniul inteligenței artificiale, deoarece reglementarea are un efect direct și imediat, și impune obligații materiale și concrete asupra furnizorilor de sisteme de inteligență artificială. Etica *recomandă comportamentul dezirabil* în timp ce reglementarea *impune comportamentul dezirabil prin forța legii.* Impoziția aceasta este legitimată de procesul deliberativ și democratic de întocmire a legii, în timp ce multe coduri de etică, care pot fi legitime - dar efectul lor este imposibil de măsurat într-un mod concret - tind să fie un amalgam ambiguu de reguli fără multă coerență și fără o reală aplicare imediată.

Pentru un viitor tehnologic mai bun - Deci, sunt roboții buni sau răi?

Se pare că răspunsul la această întrebare nu este deloc unul ușor. **Etica ne spune că roboții *pot fi buni* dacă-i concepem, de jos în sus, de la nivelul conceperii, ca fiind capabili de o etică *pozitivă*. În schimb, actualmente vedem suficiente cazuri în care roboții sunt *cam***

***răi*, dacă ne raportăm la nivelul de suferință (necuantificabilă într-un mod precis) pe care-l provoacă astăzi prin sisteme automatizate de discriminare.**

Etica în domeniul inteligenței artificiale este de o importanță absolută. Doar că, momentan, etica este folosită mai mult ca o etichetă goală, pentru marketing, imaginea de brand și pentru relațiile cu publicul ale marilor companii. Etica *reală* are o valoare *existențială imediată*. Etica ne impune reflecția asupra propriei conduite într-un mod critic și imediat. În absența eticii, viața bună, pentru noi sau colectivul nostru, n-ar fi posibilă, și drepturile omului și-ar pierde orice substanță. Pentru că și drepturile omului, la rândul lor, propun o viziune asupra comportamentului uman dezirabil, drept și just, deci propun o *etică*, una *pozitivă* axată pe protecția vieții și înflorirea ei. Viața etică ne impune să respectăm aceste drepturi fragile pe care le-am dobândit printr-un proces istoric sângeros, după cum ne amintește istoricul A.C. Grayling în cartea sa despre istoria drepturilor omului numită *Towards The Light*.

Reflecția și atitudinea etica este baza oricărei decizii morale. Dacă vrem sisteme de inteligență artificiale care servesc omului, și nu omul să le servească pe ele, trebuie să regândim infrastructura inteligenței artificiale ca fiind *dependentă de etică* în procesul său decizional, fiindcă altfel riscăm să rămânem blocați la nivelul discuțiilor despre ce este etica, ce rol are ea în cadrul inteligenței artificiale și așa mai departe fără să ajungem la esența problemei.

Unele sisteme de inteligență artificială ar trebui scoase din uz dacă în mod invariabil provoacă discriminare și suferință, mai ales cele care au un istoric rasial (cum ar fi cele de recunoaștere facială). Etica pozitivă, propusă de filosoful Luciano Floridi, ne poate deschide ochii spre construcția de sisteme de inteligență artificială care facilitează gândirea și raționamentul moral: un fel de SocrateBot care te pune să te gândești bine înainte să-ți înșeli nevasta, de exemplu. Gândirea morală ne ajută să vedem tehnologia și altfel, nu doar prin prisma profiturilor, ci și prin prisma ameliorării suferinței umane. Dacă aplicăm lentila eticii pozitive în domeniul tehnologiei, teoretizată de Floridi, e greu de spus dacă sistemele IA de azi au ameliorat în vreun fel real suferința umană. Din perspectiva eticii pozitive, dacă sistemele IA de azi n-au redus calitativ suferința umană, înseamnă că există un argument etic pentru scoaterea lor din uz, sau pentru regândirea designului lor fundamental pentru a îmbunătății viața omului.

Fiindcă, altfel, la ce bună toată tehnologia, toți chatboții, toți algoritmii din lume dacă viața nu devine mai bună, ci considerabil mai rea?