DESPRE RELEVANŢA PERSPECTIVEI KANTIENE ÎN ERA INTELIGENȚEI ARTIFICIALE GENERATIVE: PROVOCĂRI ETICE ȘI TEHNICE

IUSTINA CIOBANU (NEAGU)

Universitatea din București, Facultatea de Filosofie

On the Relevance of the Kantian Perspective in the Era of Generative Artificial Intelligence: Ethical and Technical Challenges. Since the evolution of computational and communication technologies, the world becomes a global village: society and environment are seen through digital perspectives of faster and wider industrial revolutions that become widely available and influence daily our lives. Generative AI dynamically trained and used on billions of blurred knowledge graphs' multimodal elements becomes a growing technological presence with impact. New scientific and regulatory debates address, alongside the opportunities of generative AI presence in economy and society, the challenges that could, on one side, reinvent technologies and the humanity through the Singularity hypothesis, and one the other side, open its availability for unethical, improper, or illegal use of „*black box*" learning. Within dynamically challenging tsunami of big data, information and knowledge systems, the need of contemporary wisdom captured within AI could benefit from the well-known and visionary ethical yet traditional philosophical foundations. The need to adapt the technology based on well-known and accepted logics to unsafe, opaque decisions finds solid foundations in philosophical ethics. This paper critically reviews recent scientific articles on AI challenges, the Kantian perspective, and ethics foundations that argue the need of human in the loop for the current progresses of generative AI. The role of Kantianism in the ethical perception of AI and its accent on reason and reasoning for ethics, as well as other traditional ethical theories such as theory of virtue and utilitarianism, arguably request the presence of human values at the centre of AI and justify the argument of this paper that current challenges of transparency, explainability, fairness, responsibility of AI technologies can be addressed by consistent transitions from data-centred technologies to human-centred and scientifically quantifiable solutions for the next industrial revolution(s). The paper includes case studies of ethical and common-sense challenges for the use of „*black box*" models in healthcare, paediatric robotics, also refers to other potential applications such autonomous decision making or creativity. The paper ends with arguments and open questions on opportunities and challenges for AI and philosophy scientists with a focus on ethical aspects of AI technologies.

Keywords: Kantian ethics, utilitarianism, generative artificial intelligence, black box models.

Rev. filos., **LXXI**, *3*, p. 395–406, Bucureşti, 2024 DOI: 10.59277/RF.2024.71.3.08

1. SCURTĂ INTRODUCERE

Inteligența artificială (IA) poate fi abordată ca acel fundament abstract și practic ce ilumineză traseul progresului și al evoluției în domeniul tehnologic, o abordare conceptuală ce ghidează inovația și dezvoltarea tocmai pentru a permite nu doar dezvoltarea noțiunilor teoretice, ci și implementarea de inovații tehnologice ce au plecat de la nevoia individuală și societală de sprijinire și de înlocuire a efortului uman în producția industrială cu soluții mecanizate, electrificate, automatizate, robotizate și, de curând, cibernetizate. Odată cu evoluția tehnologiilor computaționale și de comunicare, lumea devine tot mai mult un „sat global": societatea și mediul sunt privite dinspre perspective ale viziunilor digitale ce evoluează extrem de rapid; accesul și deciziile bazate pe amprentele digitale devin disponibile pe scară largă și ne influențează zilnic viața etc. Inteligența artificială (IA), cu versiunile sale generative (GenAI) antrenate dinamic și utilizate pe miliarde de informații multimodale în grafuri de cunoștințe cu organizări opace, devine o prezență tehnologică în continuă creștere cu impact nedefinit. Dezbaterile științifice și cele din domeniul reglementărilor legale sau instituționale abordează nu doar oportunitățile, ci și provocările esențiale pe care le aduce Inteligența Artificială Generativă în economie și societate. Acestea ar putea, pe de o parte, să schimbe fundamental tehnologiile existente, să reinventeze conceptul de umanitate prin prisma Singularității și, pe de altă parte, să deschidă calea către utilizări care ar putea fi considerate neetice, nepotrivite sau chiar ilegale. Actualul val tehnologic al integrărilor de sisteme de informații multimodale și cunoștințe generative, nevoia de inteligență captată în IA ar putea beneficia de unele baze etice filosofice tradiționale (vizionare). Imperativul ajustării tehnologiei fondate pe logici convenționale pentru a face față luării deciziilor bazate pe informații incerte, dificil de descifrat și care prezintă potențial etic discutabil, își găsește fundamentarea solidă într-o redescoperire, reevaluare și promovare a perspectivelor filosofiei etice, abordate inter și multidisciplinar în ceea ce privește riscul de natură tehnologică[1].

Ceea ce ne propunem în acest articol este identificarea, prin câteva exemple relevante (credem), unei serii de cerințe, de necesități și de provocări ale evoluțiilor tehnologice care ar putea beneficia de o mai solidă abordare filosofică, una fundamentală grație (mai ales) și unei perspective etice corespunzătoare. Secțiunea a doua introduce concepte și provocări ale evoluțiilor tehnologice de inteligență artificială actuale. Secțiunea a treia abordează etica kantiană, utilitarianismul și teoria aristotelică a virtuții prin prisma perspectivelor etice cerute astăzi tehnologiilor și aplicațiilor de inteligență artificială. Secțiunea a patra extinde studiile de caz cu

[1] European Commission, *European approach to artificial intelligence. The EU AI Act,* 2024.

două exemple din literatura științifică actuală care ridică problematica etică în cazul aplicațiilor de inteligență artificială în domeniile medicale de diagnoză și de asistență pediatrică. Ultima secțiune discută o serie de alternative și de necesități tehnologice și etice concluzionând cu necesitatea încorporării fundamentelor etice filosofice în proiectarea, în implementarea și în utilizarea tehnologiilor de inteligență artificială, păstrând și reflectând la prezența omului în noile sisteme automate.

2. PERSPECTIVE TEHNICE: PROVOCĂRILE CREATE DE INTELIGENȚA ARTIFICIALĂ

Rolul perspectivei de tip kantian în evaluarea etică a inteligenței artificiale, spre exemplu principiile universale ale eticii kantiene, imperativul categoric și accentul pe rațiune și raționamentul etic, precum și pe prezența valorilor umane în centrul sistemelor bazate pe cunoștințe, justifică argumentul lucrării de față, şi anume că provocările actuale referitoare la transparență, explicabilitate, corectitudine și responsabilitate ale tehnologiilor IA pot fi abordate prin tranziții consistente de la tehnologiile centrate pe date la soluții controlate epistemologic de om și cuantificabile științific în cadrul viitoarelor revoluții industriale.

Modelele IA de uz general, deși în fază incipientă, au generat un val de dezbateri și de dileme care presupun că inteligența artificială ar putea dezvolta în viitor conștiință de sine, ar putea deveni singularitate depășind capacitatea umană de (auto-)control al tehnologiei și transformându-se într-un *(f)actor de risc* pentru umanitate. Conceptul de Singularitate Tehnologică este introdus pe baza unei ipoteze neverificate (dar intens mediatizate) referitoare la tehnologii incontrolabile și ireversibile cu consecințe imprevizibile pentru civilizația umană. Cea mai răspândită interpretare sugerează că, atunci când mașinile inteligente vor depăși inteligența umană, ele se vor îmbunătăți rapid într-un ritm de neimaginat, schimbând imprevizibil și fundamental societatea. Siguranța IA reprezintă activitățile de înțelegere, prevenire și atenuare ale daunelor cauzate de IA. Aceste daune ar putea fi deliberate sau accidentale, cauzate indivizilor, grupurilor, organizațiilor, națiunilor sau la nivel global, și de multe tipuri, inclusiv, dar fără a se limita la, daune fizice, psihologice, sociale sau economice, tehnice sau militare. Reconstruind ideea centrală a criticii rațiunii pure în cadrul eticii kantiene, vom încerca să răspundem ambelor poziții, pro și contra relevanței teoretice a acestui model pentru etica IA. Articolul de față va utiliza doar exemple din bibliografia selectată, cu scopul utilizării lor ca probe într-o abordare empirică, de la cazuri particulare către o perspectivă generalizatoare mai cuprinzătoare.

3. PERSPECTIVE ETICE PENTRU INTELIGENȚA ARTIFICIALĂ

3.1. ETICA KANTIANĂ: DEONTOLOGIA, BUNUL SIMȚ, INTEGRITATEA ȘI RAȚIUNEA

Etica deontologică a lui Kant adoptă o abordare bazată pe reguli. Potrivit lui Kant, există anumite legi morale universale pe care avem datoria de a le respecta. El oferă două teste pentru a determina care sunt aceste legi. Pe baza acestor teste, Kant ar spune că furtul este întotdeauna greșit, indiferent de motivații sau consecințe. Motivația lucrării de față este astfel construită pe următoarele argumente relevante subiectului modern curent de redefinire etică și inteligență artificială: 1. principiile universale ale eticii kantiene, spre exemplu imperativul categoric, oferă un cadru de bază pentru luarea deciziilor etice în aplicațiile IA. Acest cadru ar ajuta la asigurarea coerenței și la evitarea prejudecăților subiective măcar în etapele atât de necesare și dezbătute de validare a modelelor de IA și de asigurare a securității sistemelor de IA; 2. focalizarea centrată pe om a raționamentului: Kant subliniază demnitatea umană și valoarea intrinsecă, care pot ghida dezvoltarea IA către respectarea drepturilor și a libertăților omului, chiar dacă IA în sine nu le posedă. Această abordare acordă prioritate siguranței umane în fața riscurilor potențiale ale inteligenței artificiale; 3. accentul pe rațiune și raționalitate al eticii kantiene, încurajând o analiză atentă a consecințelor acțiunilor IA, ar putea asigura un cadru de transparență și responsabilitate în dezvoltarea IA.

Este, prin urmare, nevoie de legislație specifică privind contextul tehnologic actual, atât de dinamic din prismă etică, pe lângă legile nescrise care vin din conștiința (personală, civică sau socială, comunitară)? Actualele semnale (dinspre Comunitatea Europeană, Marea Britanie, Statele Unite ale Americii, dar și de la ONU, OECD, UNESCO) ar arăta că da: spre exemplu curentele norme din UK AI Safety[2], EU AI Act[3]. Argumentarea ar trebui să privească astfel de necesități în mod neutru – de aceea vom folosi ca punct de reper referințele kantiene cu relevanță pentru tehnologiile IA actuale, dar privite etic atât în contextul Criticii Rațiunii Pure, cât și al Criticii Rațiunii Practice. Kant pune accent pe obligaţiile „din datorie", însă o mare parte din relaţiile personale, familiale, profesionale par să iasă din această sferă sau cel puțin să nu fie reprezentate adecvat de ea. Kant considera că, deşi universalizarea nu este o condiţie suficientă a judecăţilor morale, este însă o condiţie necesară: dacă o judecată morală prezintă argumente întemeiate, acele argumente sunt bune pentru orice altă împrejurare similară[4].

Evident, kantianismul nu poate fi considerat prioritar în scopul validării IA pe principii etice – în primul rând prin considerente de necontemporaneitate. Astfel, argumentele împotriva kantianismului în judecarea și validarea etică a IA pot include: antropocentrismul (abordare ce pune accentul pe importanța omului

[2] UK Government, *Policy Paper: Introducing the AI Safety Institute*, 2024.

[3] European Commission, *Shaping Europe's digital future: The AI Act*, 2024.

[4] Immanuel Kant, *Întemeierea metafizicii moravurilor*, București, Editura IRI, 1995, pp. 35–66, 81–88.

sau a ființelor umane în comparație cu alte entități sau aspecte din lume) – adoptând o perspectivă antropocentrică inspirată de kantianism, se conturează posibilitatea excluderii entităților nonumane, precum Singularitatea IA în stadiile lor avansate de dezvoltare, determinând astfel potențiala discriminare sau chiar ignorarea intereselor acestora. În contextul evaluării și al validării Inteligenței Artificiale, această perspectivă poate fi criticată pentru că ar putea minimaliza nevoile, interesele și drepturile altor forme de inteligență nonumană sau chiar ale unor entități create de om, precum sistemele avansate de IA. Cu alte cuvinte, am putea spune că prin adoptarea unei abordări antropocentrice există riscul ca deciziile etice privind IA să fie influențate în mod disproporționat de perspectiva umană, ceea ce ar putea conduce la discriminare sau la neglijarea consecințelor etice pentru acele entități. Acest aspect evidențiază complexitatea implicațiilor filosofice și etice legate de integrarea și de implementarea corectă a Inteligenței Artificiale în societate. Un alt argument ar viza domeniul limitat de aplicație: etica kantiană acordă prioritate regulilor universale, dinspre care am avea dificultăți să abordăm natura nuanțată și în evoluție rapidă a provocărilor IA. Rigiditatea imperativelor categorice ar putea împiedica adaptarea punctelor de vedere generate de etica kantiană la scenariile complexe din lumea reală actuală, centrată pe eficiență, eficacitate, performanță, utilitate, cantitate și profit. Nu în ultimul rând, trebuie subliniate dificultățile empirice și metodologice în aplicarea conceptelor kantiene la studierea IA: conceptele kantiene centrale, cum ar fi libertatea voinței și capacitatea de a acționa în conformitate cu un set de principii morale, nu sunt incluse în domeniul științific al IA, ceea ce face neclar modul de implementare a principiilor kantiene în dezvoltarea și în luarea deciziilor de către sistemele IA. Această dezbatere devine cu atât mai interesantă cu cât Kant a fost un adept al exprimării matematice formale a progreselor științifice, ceea ce ar susține utilizarea modelelor de IA ca instrumente cognitive[5]. IA dezvoltată, respectând modelul kantian propus, ar deveni explicabilă, deci deschisă spre validarea principiilor etice în mod transparent, fără opacitatea multor modele actuale de IA[6].

3.2. ALTE PERSPECTIVE ETICE: ETICA VIRTUȚII ȘI UTILITARIANISMUL

Perspectivele etice kantiene nu pot oferi o acoperire omogenă, nici măcar pragmatică, cerințelor și provocărilor tehnologiilor actuale de IA. Explozia aplicațiilor gen OpenAI GPT și DALL-E, Microsoft CoPilot sau Google Gemini pentru dezvoltarea capacităților de interacțiune și de creativitate umană (de text, imagini, înregistrări audio și video etc.) intervine cu întreaga gamă de productivitate pozitivă (gamă largă de resurse și de structuri) și negativă (*deep fakes*, halucinații, copiere ilegală etc.). Într-un astfel de context, alte teorii etice pot aduce atât

[5] Gordana Dodig-Crnkovic, *Info-Computationalism and Morphological Computing Are Models of Computation and Not Just Metaphors*. *Studies in Applied Philosophy, Epistemology and Rational Ethics – Philosophy and Theory of AI*, Springer, 2013, p. 66.

[6] Max Black, *Models and Metaphors: Studies in Language and Philosophy*, Cornell, Ithaca, 1962, p. 242.

completări, cât și perspective flexibile, adaptabile contextului modern al tehnicilor de învățare automată prin criterii și principii practice, spre exemplu prin introducerea unor criterii cantitative pentru asumarea și definirea unui cadru instituțional al responsabilității și al riscului adoptării unor astfel de tehnici în domenii de cercetare și de aplicații controversate (politică, influențare a opiniei publice, protecție socială etc.). Teoriile etice utilitariste caută esența unei acțiuni în lumina consecințelor sale, care sunt interpretate ca fiind conforme sau contrare unui standard moral. Aprecierea valorică a acțiunii se reflectă în modul în care aceasta contribuie la fericirea și la binele general, fundamentându-se pe criteriul utilității și al impactului rezultatelor asupra bunăstării generale. Consecințele cele mai relevante ar fi durerea și plăcerea. În general, teoriile utilitariste caută să minimizeze durerea și să maximizeze plăcerea. De exemplu, un utilitarist ar putea argumenta că este justificat ca o persoană săracă să fure de la o persoană bogată, deoarece banii ar cauza mai multă fericire celui sărac decât ar cauza nefericire celui bogat. În mod similar, un utilitarist ar putea argumenta că o crimă este justificată dacă victima este el însuși un criminal și astfel uciderea lui ar salva 10 vieți. Astăzi observăm că motivații utilitariste pot sta la baza algoritmilor de decizie automată în învățarea semisupervizată în care statistica voturilor pentru etichetarea unei anumite variabile de ieșire decide acțiunea viitoare: cazuri cu care astăzi suntem asaltați în sugestii pentru știri, cumpărături (Amazon, eBay), dar și de analiză a sentimentelor și chiar a tehnologiilor curente de învățare adversarială care au dus la crearea falsurilor în multimedia.

Teoria Virtuții a lui Aristotel a adoptat o abordare diferită privită din prisma aplicațiilor sale în contextul modern. În vreme ce atât etica deontologică de tip kantian, cât și utilitarismul oferă formule pentru ce am putea face, etica virtuții (ca teorie normativă ce accentuează virtuțile minții și ale caracterului deopotrivă) este mai preocupată de ce tip de persoană ar trebui să fim. Dacă o persoană virtuoasă nu ar fura într-un anumit set de circumstanțe, atunci ar fi greșit să se fure în acele circumstanțe. Conceptele de învățare automată rescrise prin prisma teoriei virtuții ar crea noi provocări și probleme tehnice și etice, generând includerea surselor și a istoricului în seturile de date de învățare. Astfel de strategii de modificare a algoritmilor de învățare ar putea deschide însă și alte dezbateri, spre exemplu în cazul greșelilor de diagnostic medicale, în care radiografiile ar trebui să conțină și datele pacientului și ale expertului medical sau în traducerea unor texte clasice sau cotidiene.

4. STUDII DE CAZ

În cele ce urmează, încercăm să exemplificăm problematica și argumentele etice pentru tehnologiile GenAI prin câteva studii de caz ale provocărilor etice și de bun simț, în introducerea modelelor opace în deciziile medicale, pentru roboții folosiți în pediatrie și alternative la exemple din robotică și creativitate în general.

4.1. EXEMPLUL 1. IA ÎN PRACTICA MEDICALĂ: CUTII NEGRE CU IMPLICAȚII POZITIVE ȘI NEGATIVE.

Algoritmii „black box" și acceptarea opacității epistemologice a acestora în aplicațiile de diagnostic și prognoză în medicina actuală sunt dezbătuți de J.M. Durán și K.R. Jongsma[7]: autorii își propun să argumenteze acceptarea algoritmilor IA pe evidențe metodologice contrar opacității lor epistemologice din prisma mecanismelor explicabilității din sistemele expert. Autorii propun menținerea acestora în aplicațiile medicale pe principii epistemologice prin considerarea condițiilor de posibilitate a performanțelor și a eficienței lor. Argumentarea pornește de la faptul că epistemologia algoritmilor este baza necesară, dar nu și suficientă, a studiilor etice ale acestora și continuă cu motivarea pe baza fiabilității computaționale a noilor modele IA, oferind contextul eficienței modelelor opace. Transparența modelelor „black box" se definește epistemologic prin existența unui grad de cunoștințe despre model (spre exemplu, descrierea unor intrări și ieșiri, dar și a valorilor acestora) prin forme relevante și contextuale care au sens pentru expertul uman. Justificarea transparenței modelului inițial devine astfel problema transparenței modelului predictor. Un model opac devine acel model imposibil de a fi epistemologic controlat de expertul uman. Modelele transparente sunt însă bazate tot pe procese opace prin transferarea verificării modelului inițial către pseudomodelul simplificat de predictor. Opacitatea metodologică este rezultatul complexității structurii și implementării algoritmilor IA datorate dimensiunii și complexității codului de programare a algoritmilor și implementării lor ca programe, și formelor abstracte și de reprezentare a informației ascunse în cod prin programarea soluției și a datelor. Soluția curentă pentru rezolvarea argumentelor morale ale utilizării modelelor opace în domeniul medical este construită de autori pe fiabilitatea computațională, care oferă o justificare epistemică a încrederii în fiabilitatea algoritmului, și eliminarea nevoii utilizării predictorilor de interpretare. Mecanismul propus se bazează pe faptul că probabilitatea ca următorul set de rezultate al unui sistem de IA fiabil să fie de încredere este mai mare decât probabilitatea obținerii unui set de rezultate de încredere produse la întâmplare. Acest mecanism de adoptare de tehnologie IA pare a fi motivat de încrederea validată empiric. Dar cum putem avea încredere în primele rezultate? În IA, acestea se obțin prin procese de verificare și de validare cu date experimentale ale robusteții și acurateței modelelor considerate sau prin simulări computaționale. Odată validate, aceste modele continuă să fie folosite în sisteme de decizie medicală, iar importanța lor este susținută de calitatea datelor clinice, de interpretarea rezultatelor în contextul menținerii expertului uman în proces și de prioritizarea siguranței pacientului.

[7] Juan Manuel Durán, Karin Rolanda Jongsma, „Who is afraid of black box algorithms: epistemological opacity vs methodological transparency", în *Journal of Medical Ethics*, BMJ, 2021, pp. 329–335.

Autorii argumentează și lipsa de opțiuni pentru pacienți, deoarece prioritățile pacienților pot fi diferite de cele ale modelelor IA. De aceea este necesară includerea expertului și a pacientului uman în ciclul de decizie medicală, mai ales pentru sistemele de IA. Concluzia autorilor este că utilizarea algoritmilor IA în practica medicală necesită încrederea epistemică în algoritm și în rezultatele sale, dar aceasta nu este și suficientă pentru integrarea lor. Medicii pot avea încredere în algoritmi IA „black box" care sunt fiabili epistemic, chiar justificați normativ, pe baza evidențelor. O atitudine clar utilitaristă, care însă cu greu se poate justifica dinspre un orizont kantian sau aristotelician.

Etica IA se bazează pe epistemologia algoritmilor, dar nu justifică integrarea independentă, fără expertul și utilizatorul uman, în procesul de decizie. Consecințele morale ale deciziei cu modele IA necesită ca astfel de modele să fie interpretabile și transparente. Percepția publicului este încă împărțită în ceea ce privește utilizarea algoritmilor fără integrarea contextului medical al pacientului și al factorilor de responsabilitate în luarea deciziei. Pașii următori necesită evidențe cu valoare științifică, antrenarea personalului medical și includerea utilizatorilor umani (personal medical, pacienți, experți legali sau etici) în procese de decizie și de tratament medical. Această concluzie este dificil de acomodat (în practică) cu teoria deontologică a lui Kant.

4.2. EXEMPLUL 2: SISTEME IA ÎN APLICAȚII ROBOTICE PEDIATRICE

Un studiu de caz al roboților folosiți în domenii medicale cu sensibilitate crescută, spre exemplu în domeniul pediatric, este descris de J. Borenstein, A. Howard și AR Wagner[8]. Cum putem construi și valida încrederea în roboți inteligenți care ar putea rezolva eficient problema cozilor de așteptare (interminabile) în spitale, în camera de triaj la urgențe sau în spitalizare atunci când este în joc siguranța copiilor, mai ales al celor cu comportamente cognitive sau mobilități anormale? Care este consecința interacțiunilor copil-robot în medii nesupravegheate? Ar deveni copiii mai puțin comunicativi, ar dezvolta comportament antisocial sau asocial? Continuând pe aceeași temă, autorii explorează potențialul necunoscut al includerii roboților inteligenți în mediul de acasă. Argumentul principal în astfel de cazuri este concentrat pe atitudinea de supraîncredere în tehnologie a individului, fie copil, fie părinte sau chiar familie. Evident, măsuri de evitare a unor astfel de dileme și de dificultăți pot fi proiectate rezonabil de repede, dar aducând în plus propriile dificultăți: spre exemplu, robotul va interacționa cu copilul doar în prezența (fizică sau virtuală) a unui părinte sau adult aprobat de sistem… dar este această condiție realistă? Metodologiile de confirmare etică și de siguranță includ astfel actualele cerințe ale omului în ciclul de funcționare a IA în medii delicate social din

[8] Jason Borenstein, Ayanna Howard, Alan R. Wagner, „Pediatric Robotics and Ethics: The Robot Is Ready to See You Now, But Should It Be Trusted?", în Patrick Lin, Keith Abney, Ryan Jenkins, *Robot Ethics 2.0*, Oxford University Press, 2017, pp 127–141.

punct de vedere al responsabilității. Legislația, regulamentele și instrucțiunile de asigurare vor face astfel parte din contextul extinderii roboților în aceste medii personalizate, spre exemplu prin considerarea includerii eticii kantiene încă de la nivelul proiectării unor astfel de roboți – devenind practic sisteme etice (kantiene) prin proiectare.

5. DISCUȚIE ȘI CONCLUZII

Care ar fi comportamentul unui „robot kantian”? Dar al unui „robot virtuos” sau al unui „robot utilitarist”? În primul caz, va fi probabil un adept și un exemplu al imperativului categoric, inclusiv al principiului respectului pentru oameni, chiar dacă noi nu suntem exemple sau adepți ai eticii kantiene. Este atunci prudent sau justificat să le cerem roboților să demonstreze asemenea tipuri de comportamente? Evident, realizarea și impactul unor astfel de obiective sunt încă incerte[9].

Centrarea articolului de față pe abordarea kantiană se datorează în primul rând capacității și extinderii argumentului uman și limitărilor aproape imposibil de implementat de tehnologiile actuale – prin urmare oferind șansa alternativei filosofice ca termen de referință optim și idealist. Teoriile curente de GenIA și idealurile tehnologice de singularitate ce par a fi guvernate de performanță și de cantitate nu au avut nicio șansă să introducă principii etice în algoritmii de învățare automată. Focalizarea teoriilor etice, în general, și a filosofiei kantiene, în special, pe norme universale care nu sunt dependente sau influențate de context par a fi nu doar problematice pentru om ca individ, dar și pentru actualele soluții tehnologice de IA centrate pe date și pe contextul de culegere a datelor de antrenare, testare și validare și pentru gestionarea acestora pe criterii de calitate măsurabile în general *a posteriori*.

5.1. CUM SE INCORPOREAZĂ CONȘTIINȚA ȘI BUNUL SIMȚ ÎN IA?

Desigur că ideea de implementare a raționamentului moral încă din stadiul de proiectare a IA – spre exemplu bazat pe principiile eticii kantiene (printre altele) – nu este una nouă. Asemenea tipuri de proiecte IA ar urmări să implice și să includă în modelele de inteligență computațională gramatici morale, sisteme bazate pe reguli ale principiilor deontologice sau, dimpotrivă, ale principiilor utilitariene sau soluții hibride. Michał Klincewicz propune, printre alte direcții esențiale, investigarea algoritmilor pentru imperativul moral fundamentat pe deontologia kantiană[10]. Autorul examinează provocările întâmpinate în implementarea deontologiei

[9] Thomas M. Powers, „Prospects for a Kantian Machine”, în *IEEE Intelligent Systems*, vol. 21, no. 4, 2006, pp. 46–51.

[10] Michał Klincewicz, „Challenges to Engineering Moral Reasoners: Time and Context”, în Patrick Lin, Keith Abney, Ryan Jenkins, *Robot Ethics 2.0*, Oxford University Press, 2017, pp. 244–257.

kantiene (și a altor teorii etice, inclusiv utilitarismul, care se confruntă cu probleme similare) în algoritmii inteligenței computaționale. Argumentul central se referă la ideea că teoria deontologică, în formularea ei kantiană, susține că moralitatea derivă din intenția sau din voința unei entități umane responsabile de o stare mentală. Imperativul categoric este perceput ca o lege morală fundamentală a rațiunii umane: astfel încât, dacă intenția urmează acest imperativ, atunci acțiunea care decurge este considerată morală; în schimb, dacă intenția nu respectă imperativul categoric, atunci aceasta este considerată imorală. Având în vedere faptul că tehnologiile computaționale actuale nu sunt capabile să demonstreze stări mentale (cel puțin încă nu există dovezi ale atingerii unei stări de singularitate a IA), ipoteza existenței unui program cu o rațiune morală kantiană rămâne incertă. Conceptul de singularitate IA, numit program cu rațiune morală kantiană de către autor, reprezintă un program cu stări mentale ce ar evidenția intenții și raționament la nivel uman – nivel neatins (încă?) din punct de vedere tehnologic. Controversele suscitate (publicate recent) de soluțiile care folosesc IA de tip generativ (ChatGPT, Bard, Copilot, Genesis etc.) pleacă de la ipoteza că acest tip de tehnologii demonstrează comportament uman grație răspunsurilor de calitate creativă (bazate pe numărul imens de variabile și date de antrenare) și a capacității de comunicare cu utilizatorii umani folosind semnalele relevante comunicării între oameni (text, imagini, voce/sunet, video). Există deja dovezi că aceste IA generative sunt capabile să treacă teste de admitere la facultăți de medicină, de psihologie sau de inginerie și demonstrează abilități de comunicare și inferențe capabile să treacă anumite variante ale testului lui Alan Turing. Argumentul autorului trece însă de imposibilitatea creării stărilor mentale computaționale și argumentează prin divizarea problemei imediate complexe în cele trei formulări componente. Autorul presupune posibilitatea implementării prin tehnologii curente de IA a algoritmilor imperativelor morale kantiene care se bazează pe imperativul categoric din teoria etică kantiană, exprimată prin cele trei formulări.

Într-adevăr, problematica implementării raționamentului kantian se mută în discursul autorului de la potențialul de implementare a stărilor mentale umane la validarea triplei formulări a maximei universale. Cu alte cuvinte, includerea validării automate a unor stări de fapte generalizate maximale, de genul „nu mai există pâine pentru nimeni”, necesită încrederea în senzorii, în intrările și în comunicările sistemului IA cu lumea exterioară. Autorul explorează posibilitățile de rezolvare pornind de la teoriile actuale de ingineri a cunoștințelor pentru examinarea posibilelor contradicții asociate cu raționamentele menționate anterior. Se aduc în discuție opțiuni interesante, precum sistemele expert și sistemele bazate pe cunoștințe. Argumentul se extinde cu referiri la reprezentarea cunoștințelor în format XML, deși se constată că astfel de abordări sunt, în prezent, marginale în cadrul sistemelor IA generative care recurg la date sintetice din lipsa datelor concrete. În final, se conchide că sistemele care încorporează moralitatea kantiană ar putea beneficia de un avantaj tehnologic comparativ cu cele utilitariste, datorită modalității de reprezentare a cunoștințelor care se bazează pe informații pure, fără reguli explicite, însă care, la rândul lor, necesită evaluare din perspectivă etică.

În sprijinul acestei linii de argumentare, Fabio Bonsignorio abordează utilitatea ideilor lui Kant referitoare la sine și la conștiință în contextul actual[11]. Metoda trascendentală kantiană este corelată cu tendințele și cu cerințele actuale din domeniul științelor cognitive: pentru a examina inteligența, trebuie să confirmăm mecanismele mentale ale subiectului experimentului, prin integrarea coerentă a tuturor experiențelor sale. Această perspectivă oferă baze epistemologice pentru argumentele științifice din discuțiile și dilemele contemporane legate de posibilele riscuri asociate dezvoltării viitoare a noilor sisteme de IA ca entități tehnologice autonome și de validarea acestora. Fabio Bonsignorio intensifică astfel standardele în ceea ce privește validarea etică și securitatea tehnică și legislativă a sistemelor de IA, subliniind importanța analizelor și a investigațiilor filosofice propuse de filosofi precum Kant, Hume și Locke (printre alții) în contextul studiilor contemporane asupra eticii în domeniul Inteligenței Artificiale.

Așa cum aceste studii, în trecut, au facilitat progresul de la studierea naturii la științele moderne, tot astfel investigațiile filosofice asupra minții și a inteligenței umane ar putea influența progresul tehnologic curent în IA și în etică. Necesitatea acestui tip de cerințe este chiar principial subliniată de către autor care, într-o manieră ironică, dar susținută în mod consecvent cu referințe filosofice, afirmă că „vechii filosofi", printre care și Kant, au fost mult mai atenți la potențialele erori logice rezultate din lipsa cunoașterii decât inginerii și proiectanții contemporani de sisteme IA, roboți și sisteme cibernetice (cum ar fi sistemele IA de la BostonDynamics, OpenAI, Anthropic) – astfel de entități fiind, prin natura lor, opace din cauza lipsei de transparență a resurselor digitale pe care se bazează modelele lor generative de învățare.

5.2. CONCLUZII

Dezvoltările unor tehnici și domenii noi care includ Inteligența Artificială bazate pe revoluțiile industriale tehnologice, dar și evoluțiile profesionale și mediatice sunt doar un alt pas în evoluția societății umane. În drumul nostru către o comunitate de producători și de utilizatori responsabili de tehnologie, proiectanții (umani sau artificiali) au nevoie de includerea sustenabilă și fiabilă a conceptelor filosofice relevante, printre care credem că un model kantian etic-epistemologic poate contribui în mod cert la validarea acestora. Rolul acestui model în configurarea unei perspective „etice" asupra sistemelor IA este crucial, putând furniza principii esențiale pentru abordările actuale referitoare la emanciparea și la personalizarea

[11] Fabio Bonsignorio, „Section 2.10 The New Experimental Science of Physical Cognitive Systems: AI, Robotics, Neuroscience and Cognitive Sciences under a New Name with the Old Philosophical Problems?", în *Studies in Applied Philosophy, Epistemology and Rational Ethics – Philosophy and Theory of AI*, Springer, 2013, pp. 139–145.

tehnologiilor IA prin argumentare rațională. Chiar dacă există posibile limite determinate de nesincronizări temporale, tehnologice sau sociale care pot necesita o analiză mai atentă, o considerare nuanțată și potențiale referiri la alte cadre etice, această dezbatere rămâne deschisă, necesitând discuții continue, deoarece tehnologiile IA continuă să evolueze.