INTELIGENȚĂ ARTIFICIALĂ MORALĂ? DESPRE ALGORITMI, VIRTUȚI ȘI STATUT MORAL

Autor: Mihaela Constantinescu | 30 ianuarie 2022

Putem codifica moralitatea într-un sistem de Inteligență Artificială, care să devină astfel capabil să ia decizii morale și, pe deasupra, să facă acest lucru independent de intervenția umană? Iată întrebarea care a generat o întreagă direcție de cercetare în cadrul mai larg al studiului Inteligenței Artificiale (IA), la intersecția dintre filosofie, informatică, științe cognitive și psihologie, denumită generic etica mașinilor (*machine ethics*), alternativ moralitate artificială (*artificial morality*). Asumpția acestei direcții de cercetare este că statutul moral pe care îl asociem în prezent exclusiv oamenilor, acela de ființe capabile să acționeze și să delibereze moral (să fie, altfel spus, agenți morali), poate fi extins, de principiu, și altor entități. Astfel că scopul cercetărilor privind moralitatea artificială este acela de a crea astfel de entități, mai precis Agenți Morali Artificiali Autonomi (*Autonomous Artificial Moral Agents*). Mai mult sau mai puțin inspirați de scrierile lui Asimov, cercetătorii care propun în secolul XXI variante de a construi Agenți Morali Artificiali Autonomi au în vedere mai cu seamă algoritmi bazați pe învățare automată (*machine learning*), care folosesc rețele neuronale ce permit învățarea stratificată (*deep learning*). Acești algoritmi pot deveni, ulterior, parte a unui sistem robotic, cum este cazul autovehiculelor fără șofer, al roboților companioni sau al unor roboți-educatori.

Dar este posibil să generăm astfel de entități morale artificiale, cu un grad ridicat de autonomie acțională și deliberativă? Am pornit în căutarea unui răspuns la această întrebare din perspectiva eticii virtuții aristotelice[1] alături de colegii mei, Cristina Voinea, Radu Uszkai și Constantin Vică, în cadrul proiectului de cercetare CoMoRe[2], demarat la începutul anului 2021. Concluziile la care am ajuns până în prezent îi vor liniști pe unii și, probabil, îi vor tulbura pe alții.

Pentru a putea vorbi de Agenți Morali Artificiali Autonomi, ar trebui să avem mai întâi o definiție, cel puțin de lucru, pentru ce înseamnă un agent moral. Ei bine, o astfel de definiție nu există și nu avem nici măcar un consens cu privire la setul de condiții

necesare (cu atât mai puțin cu privire la setul de condiții suficiente) pe care o entitate ar trebui să le îndeplinească pentru a dobândi statutul de agent moral. Sunt propuse, printre altele, aspecte precum intenționalitate, autonomia voinței și deliberării, capacitatea de a simți, de a înțelege semnificația morală a faptelor, conștiența, auto-reflexivitatea[3]. În prezent, ființele umane adulte reprezintă singura categorie pentru care există un consens cu privire la capacitatea de a primi statutul de agenți morali. În timp ce unii cercetători mai optimiști sunt încrezători că sistemele de IA pot dobândi, într-un viitor mai apropiat sau mai îndepărtat, acest statut, cei mai mulți rămân sceptici cu privire la această posibilitate. Aceștia din urmă aduc în discuție aspecte limitative precum dependența sistemelor de IA de creatorii lor umani, incapacitatea de a înțelege și interpreta valorile umane și perspectivele morale ale comunităților umane, ori capacitatea de a simți emoții sau de a simți în general. De cealaltă parte, cercetătorii optimiști văd în algoritmii ce operează pe baza învățării automate un potențial agent moral artificial, realizabil în viitorul nu foarte îndepărtat, ca sistem capabil să învețe comportamente morale și să le pună în act întocmai ca o ființă umană. Un astfel de agent moral artificial, susțin aceștia, ar putea fi educat, de exemplu, să dobândească virtuți etice asemenea unui copil[4]. Dar este oare posibil?

Folosind cadrul eticii virtuții aristotelice, considerată în prezent una dintre cele mai influente trei teroii etice (alături de utilitarism și deontologism kantian)[5], răspunsul este mai degrabă nu. Nu putem concepe algoritmi sau sisteme artificiale morale virtuoase, cel puțin nu în prezent sau în viitorul apropiat[6]. Oricât am încerca să educăm virtuți morale în rândul algoritmilor, cele mai avansate mecanisme de învățare automată stratificată pe care se bazează aceștia nu fac decât să genereze o similitudine comportamentală cu agenții umani virtuoși. Altfel spus, sistemul de IA va arăta că poate identifica și pune în act comportamente virtuoase, dar nu va fi un agent moral virtuos, cel puțin nu în sensul în care atribuim acest statut moral unei ființe umane. În parte, pentru că sistemul de IA nu își poate fixa singur scopurile, ceea ce îl face ne-autonom, dependent de scopurile programatorilor, designerilor și utilizatorilor. În terminologie aristotelică, *telos*-ul sistemelor de IA este așadar fixat extern. Apoi, pentru că sistemele de IA care „învață” despre virtuți nu pot dezvolta motivațiile morale cerute de cadrul eticii aristotelice: nu este suficient ca sistemele de IA să acționeze virtuos, ci mai este nevoie și ca acțiunea lor să fie bazată pe motivațiile adecvate și să fie adaptată la contextul situațional specific. Or, sistemele de IA nu au capacitatea de a dezvolta motivațiile interne potrivite și nici nu au potențialul de a deveni un *phronimos* aristotelic, de a intui acțiunea virtuoasă relativ la variațiile contextului. Pentru aceasta sunt necesare, în perspectiva aristotelică, afectele, capacitatea de a simți, parte inerentă a biologiei umane. Iar tot acest proces este mult mai complex decât simpla aplicare mecanică a unor reguli generale în diferite contexte particulare[7]. Sistemele de IA antrenate pe seturi vaste de date din care învață în mod automat despre comportamente virtuoase nu au însă experiența practică necesară, istoria personală și nici motivațiile intrinseci necesare pentru un comportament virtuos autentic[8].

Nu putem vorbi, așadar, de Inteligență Artificială morală, cel puțin nu din perspectiva eticii virtuții: nu putem construi Agenți Morali Artificiali Autonomi care să fie realmente virtuoși. Putem, în schimb, să utilizăm cadrul eticii virtuții pentru a înțelege și evalua implicațiile etice ale dezvoltării și utilizării sistemelor de IA. De exemplu, pentru a analiza în ce măsură este de dorit să proiectăm și să utilizăm cu toată încrederea sisteme inteligente care să fie puse în situația de a lua decizii morale, fie că vorbim de autovehicule fără șofer care pot vătăma pietoni pentru a a-și salva pasagerii, de algoritmi care dau verdicte cu privire la eliberarea condiționată a unor deținuți sau de roboți umanoizi proiectați pentru a deveni cei mai buni prieteni ai copiilor. Dincolo de aceste exemple problematice, cadrul eticii virtuții ne ajută să fim atenți și la situații aparent mai puțin complicate, precum cea în care utilizăm sistemele de IA sub forma unui asistent moral, o aplicație pe care o utilizăm, de pildă, pentru a consulta mai multe repere morale atunci când avem de luat decizii cu implicații etice. În funcție de modul în care a fost proiectată și de gradul în care ne bazăm pe o astfel de aplicație, poate apărea un risc de altă natură, asupra căruia atrage atenția Shannon Vallor[9]: acela de a ne pierde, în timp, (o parte din) abilitățile morale, din înțelepciunea practică sau flerul nostru moral.

Există însă și utilizări fericite, chiar multe, ale algoritmilor de IA în contexte cu încărcătură morală. Un exemplu sunt aplicațiile de *co-parenting*[10], folosite deja cu succes în S.U.A. sau Marea Britanie de mulți părinți separați, pentru a gestiona mai bine comunicarea dintre ei atunci când planifică programul copiilor în co-tutelă. Astfel de aplicații folosesc algoritmi de analiză a sentimentelor (nu lipsiți de controverse, la rândul lor) și îi semnalează utilizatorului dacă mesajul pe care urmează să îl trimită conține expresii ce pot fi interpretate drept ofensatoare sau agresive și îi dau astfel posibilitatea să modifice mesajul respectiv înainte de a-l trimite.

Esențială rămâne, în toate aceste contexte, înțelepciunea umană: înțelepciunea programatorilor și dezvoltatorilor de a înțelege implicațiile morale ale aplicațiilor algoritmilor în proiectarea și implementarea sistemelor de IA, înțelepciunea designerilor de a evalua corect impactul aspectului roboților asupra utilizatorilor, înțelepciunea reglementatorilor în a legifera cât este necesar, când este necesar. Și poate, mai ales, înțelepciunea utilizatorilor, categorie în care ne aflăm cei mai mulți dintre noi, de a nu vedea în aceste sisteme mai mult decât sunt sau ar trebui să fie: tehnologii menite să ajute, să sprijine, nu să dezbine sau să distrugă.

NOTE

1. Așa cum reiese cu precădere din *Etica Nicomahică* a lui Aristotel. Recomandare de lectură în limba engleză: Aristotle. (2018). *Nicomachean Ethics.* Second edition (trans

and ed: Crisp, R.). Cambridge: Cambridge University Press. ↑

2. Proiectul *CoMoRe – Responsabilitate morală colectivă: de la organizații la sisteme artificiale. O re-evaluare a cadrului aristotelic* este finanțat printr-un grant al Unității Executive pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării, CNCS–UEFISCDI, cod PN-III-P1-1.1-TE-2019-1765 și implementat în cadrul Centrului de Cercetare în Etică Aplicată (CCEA) și ICUB, Universitatea din București. ↑

3. A se vedea propuneri precum: i) Hakli, R., & Mäkelä, P. (2019). Moral responsibility of robots and hybrid agents. *The Monist, 102*(2), 259–275; ii) Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology, 16*(3), 197–206; iii) Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29; iv) Misselhorn, C. (2018). Artificial morality. Concepts issues and challenges. *Society, 55*(2), 161–169; e) Véliz, C. (2021). Moral zombies: Why algorithms are not moral agents. *AI & Society, 36*, 487–497. ↑

4. De exemplu, Howard, D. & Muntean., I. (2016). A Minimalist Model of the Artificial Autonomous Moral Agent (AAMA). *AAAI Spring Symposia*; Wallach, W. & Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press sau Malle, B.F. (2016). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 18, 243–256. ↑

5. Pentru o analiză pertinentă care pune în relație cele trei teorii etice recomand cartea lui Cristian Iftode, *Viața bună. O introducere în etică*, apărută în 2021 la editura Trei. ↑

6. Argumentarea pe larg a acestui punct de vedere se găsește în articolul „Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context”, *Ethics and Information Technology*, 30 septembrie 2021, în acces liber: https://link.springer.com/article/10.1007/s10676-021-09616-9. ↑

7. Irwin dezvoltă acest punct de vedere în introducerea pe care o face traducerii textului aristotelic: Aristotle. (1999). *Nicomachean Ethics* (trans. and ed. T. Irwin), second edition (pp. xiii-xxviii). Indianapolis: Hackett Publishing Company, Inc. ↑

8. Punct de vedere dezvoltat în Sparrow, R. (2021). Why machines cannot be moral. *AI & Society*, https://doi.org/10.1007/s00146-020-01132-6. ↑

9. A se vedea Vallor, S. (2015). Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. In *Philosophy & Technology*, *28*, 107-124. ↑

10. Precum OurFamilyWizard, coParenter, TalkingParents – o analiză pe larg în Coldwell, W. (2021, December 29). What Happens When an AI Knows How You Feel? Technology used to only deliver our messages. Now it wants to write them for us by understanding our emotions. *Wired*.

https://www.wired.com/story/artificial-emotional-intelligence/. ↑

Imagine: *Unsplash*