INTERVIU CU DOAMNA PROFESOARĂ MIHAELA CONSTANTINESCU: „ÎNȚELEPCIUNEA PRACTICĂ ESTE CEVA CE SISTEMELE DE IA NU POT AJUNGE SĂ DEZVOLTE”

Autor: Victoria Maria Deliu | 31 iulie 2022

Mihaela Constantinescu *este lector la Facultatea de Filosofie și director executiv al Centrului de Cercetare în Etică Aplicată (CCEA), Universitatea din București. În perioada 2021-2022 coordonează proiectul național de cercetare CoMoRe privind responsabilitatea morală asociată dezvoltării sistemelor de inteligență artificială și analizată din perspectiva eticii virtuții aristotelice. Interesele sale de cercetare includ etica afacerilor, etica inteligenței artificiale și etica virtuții. A publicat articole pe teme ce țin de responsabilitate morală individuală și colectivă, infrastructuri etice, statutul moral al sistemelor de inteligență artificială sau interacțiunea om-robot în reviste academice precum Business Ethics, the Environment and Responsibility, Journal of Business Ethics, Philosophy & Technology, Frontiers in Robotics and AI, etc. Semnează, alături de Valentin Mureșan, volumul Instituționalizarea eticii: mecanisme și instrumente. Este pasionată de etica pentru copii și dezvoltă ateliere non-formale de gândire critică morală pentru copii și adolescenți în cadrul Asociației pentru Educație în Științe Socio-Umane (ESSU).*

1. În cadrul Centrului de cercetare în etică aplicată (CCEA), afiliat Facultății de Filozofie a Universității din București, coordonați, începând cu ianuarie 2021 și până în decembrie 2022, proiectul de cercetare *CoMoRe* sau *Responsabilitatea morală colectivă: de la organizații la sistemele de inteligență artificială. O*

***reevaluare a eticii virtuții în tradiție aristotelică*. Pornind de la întrebarea „Cine este responsabil din punct de vedere moral pentru deciziile luate de sistemele de inteligență artificială, în special de cele care sunt în mare măsură autonome?”, ați discutat, alături de colegii dumneavoastră, în articolele publicate până acum în cadrul acestui proiect, despre responsabilitatea morală și posibilitatea de a o atribui sistemelor de inteligență artificială, respectiv creatorilor lor. Ne-ați putea spune pe scurt care sunt concluziile la care ați ajuns până în acest moment în acest studiu și ce urmează să tratați în continuare?**

În ultimii câțiva ani s-a vorbit tot mai des despre utilizarea inteligenței artificiale (IA) pentru sisteme precum mașini autonome, drone autonome sau arme autonome, dar și pentru sisteme de recunoaștere facială, recrutare sau acordare de împrumuturi bancare. Toate acestea sunt exemple în care IA este dezvoltată pentru a completa sau chiar înlocui oamenii în diverse procese de luare a deciziilor, lucru valabil mai ales în cazul IA bazate pe învățare automată (mai cu seamă învățare automată stratificată – *deep learning*) și a cărei funcționare este adesea opacă, neinteligibilă nu doar pentru utilizatori, ci și pentru programatori. Am avut și suficiente exemple despre efectele negative ale automatizării deciziilor: accidente în care mașinile autonome au ucis oameni în faza de testare, sisteme de recunoaștere care nu pot identifica persoanele afro-americane, recrutarea bărbaților în detrimentul femeilor doar pe criterii de gen și lista poate continua.

Întrebarea căreia ne-am străduit să îi oferim un răspuns în cadrul proiectului CoMoRe este: cine poate fi considerat responsabil din punct de vedere moral pentru deciziile luate de astfel de sisteme autonome, care nu execută simple comenzi, ci funcționează pe baza unor algoritmi ce învață din experiență și generează rezultate pe care nici echipa de programatori, ingineri și dezvoltatori nu le poate anticipa? Deși sunt voci în comunitatea de cercetare care sunt gata să acorde statutul de agent moral responsabil acestor sisteme, concluziile principale la care am ajuns până acum în cadrul CoMoRe arată că sistemele artificiale autonome din prezent și pe care le putem anticipa tehnologic nu justifică acordarea unui astfel de statut, cel puțin nu din perspectiva cadrului eticii virtuții aristotelice pe care l-am folosit. În acest cadru de evaluare, pentru a putea fi un agent moral, așadar un agent care să ia decizii morale pentru care să fie responsabil, un sistem bazat pe IA ar trebui să poată acționa virtuos, să dezvolte atât virtuți de caracter (etice) cât și intelectuale (dianoetice). Chiar dacă putem fi de acord că algoritmii unei mașini fără șofer sau ai unui joc complex precum AlphaZero procesează date, realizează inferențe, stabilesc conexiuni complexe, iau decizii raționale, astfel că pot ajunge să dezvolte anumite virtuți dianoetice (precum *nous* sau *episteme*) corelate cu studiul teoretic, nu același lucru se poate spune despre virtuțile etice. Acestea din urmă pot fi dezvoltate doar prin învățare participativă și necesită un anumit tip de motivație, dar mai ales înțelepciune practică (*phronesis*). Or, înțelepciunea

practică este ceva ce sistemele de IA nu pot ajunge să dezvolte pentru că nu pot avea acea înțelegere a elementelor contextuale relevante pentru decizia morală virtuoasă. Pe scurt, sistemele de IA nu pot fi agenți morali responsabili pentru că le lipsește, printre altele, înțelepciunea practică, nu pot fi un *phronimos* aristotelic. Astfel, responsabilitatea morală pentru deciziile luate și efectele produse de sistemele artificiale le revine exclusiv oamenilor implicați în adoptarea și punerea în practică a acestor decizii.

Următoarea provocare este de a privi atent la această rețea complexă de oameni și responsabilități individuale asociate implementării sistemelor artificiale autonome, dar și la un nivel colectiv de responsabilitate morală din care fac parte organizații și instituții sociale sau politice. Dar sunt și alte direcții tematice conexe la care am început să lucrăm și pe care m-aș bucura să le dezvoltăm și după terminarea proiectului la finalul acestui an. CoMoRe este pentru mine mai mult decât un proiect de cercetare, este un proiect de suflet în care am avut bucuria să îi am alături pe Cristina Voinea, Radu Uszkai și Constantin Vică, toți colegi la CCEA. Unul dintre articolele care urmează să fie publicat în scurt timp tratează tema prieteniei dintre copii și roboți, în care păstrăm cadrul eticii virtuții aristotelice pentru a susține că nu putem vorbi de prietenie-virtute, prietenie perfectă între copii și roboți, cel puțin nu în sensul pe care îl avea în vedere Aristotel. În schimb, susținem că folosirea roboților companioni dotați cu inteligență artificială, segment care s-a dezvoltat foarte mult în ultimii doi ani în contextul izolării sociale, poate contribui la formarea virtuoasă a copiilor. Aș sublinia aici că, din nou, excludem posibilitatea ca roboții companioni să fie considerați ei înșiși entități morale virtuoase în relaționarea cu copiii, nu sunt un model la care copiii să se raporteze. Roboții care respectă o serie de cerințe – despre care vorbim pe larg în articol – pot fi însă companioni de joacă ce îi ajută pe copii să dezvolte comportamente virtuoase față de oamenii din jur.

2. Date fiind aceste concluzii – anume că un sistem de inteligență artificială, deși poate avea anumite virtuți dianoetice, rămâne lipsit de înțelepciunea practică și nu se poate, așadar, apropia de statutul de *phronimos* –, responsabilitatea morală pentru deciziile luate de sistemele de inteligență artificială ar trebui atribuită mai degrabă programatorilor sau creatorilor acestor sisteme. Ce provocări credeți că întâmpină astăzi programatorii, în dezvoltarea unor sisteme de IA în acord etica virtuții, dacă ne gândim că cei mai mulți dintre ei lucrează în corporații, care par să promoveze valori destul de diferite de virtuțile clasice? Și cum pot fi depășite aceste provocări?

Nu pot spune că am un răspuns simplu și direct la aceste întrebări, deși mi-ar fi plăcut să pot oferi unul. Responsabilitatea morală pentru efectele sistemelor de IA are o

dimensiune pronunțat colectivă și este mai degrabă o rețea complexă de oameni, organizații și instituții, cu multe relații de interdependență între componente. Ar fi ușor să dăm vina doar pe programatori, dezvoltatori, ingineri, designeri sau chiar utilizatori, dar nu ar fi suficient. Programatorii, inginerii și designerii sistemelor de IA sunt parte din diferite organizații, unele start-up-uri, altele centre de cercetare, altele companii multinaționale, fiecare cu o cultură organizațională și anumiți indicatori de performanță asumați. Sunt zone în care oamenii care proiectează sistemele de IA în cadrul acestor organizații au un cuvânt de spus și zone în care sunt culturi osificate, cu practici greu de transformat în mai bine. Din acest motiv cred că trebuie să includem și entități de tipul organizațiilor în rețeaua complexă de responsabilitate morală în jurul IA, tocmai pentru că sunt situații în care nu doar anumiți indivizi, ci și organizații întregi trebuie să fie blamate.

Este important ca oamenii din zona tehnică să conștientizeze rolul pe care îl au și implicațiile multiple ale utilizării sistemelor de IA pe care le concep, iar acest lucru poate fi obținut prin educație la diferite niveluri. În Țările Scandinave, mai ales Suedia, se discută deja serios despre un curriculum universitar trans-disciplinar și introducerea unor cursuri din sfera socio-umanelor în universitățile tehnice. Probabil acest tip de educație morală în direcția dezvoltării și utilizării etice a IA ar trebui să înceapă chiar mai devreme, la nivel gimnazial. Cei care sunt deja activi în zona tehnică de dezvoltare IA pot beneficia de o înțelegere mai amplă a efectelor sociale ale IA prin training-uri etice specializate de gândire critică morală, prin care să dobândească inclusiv abilitatea de a identifica situațiile problematice moral (chiar dacă nu știu cum să le rezolve). Organizațiile pot apela în plus la centre independente de eticieni – cel mai adesea vedem că sunt angajați experți etici dar și aceștia vor avea de respectat, la rândul lor, o serie de constrângeri interne ale organizației respective.

Nu în ultimul rând, reglementarea prin legislație primară și prin recomandări, în special prin standarde și certificări, poate contribui substanțial la îmbunătățirea practicilor din domeniu. Avem în prezent seria IEEE P70xx de standarde de etică și certificarea asociată pentru sisteme inteligente și autonome, în completarea ghidurilor și recomandărilor elaborate de organisme precum UE, OECD, UNESCO. Suntem însă încă în etapa vestului sălbatic, în lipsa unei legislații adecvate sau măcar a presiunii publice ca organizațiile să folosească certificări de tipul IEEE.

3. Păstrând în minte tradiția morală aristotelică și virtuțile clasice – precum curajul, cumpătarea, prudența și dreptatea – de ce considerați că astăzi, sub eticheta „virtuți” trec mai frecvent deschiderea, acceptarea, toleranța decât virtuțile despre care vorbea Aristotel?

E interesant că virtuțile se schimbă în ritm cu epocile, în timp ce sufletul omului, în care

sunt cultivate virtuțile, probabil că nu a suferit mutații substanțiale din vremea în care scria Aristotel. Condițiile exterioare și raportarea la cei din jur se schimbă însă, ceea ce ne permite astăzi să vorbim de etica virtuții în tradiție neo-aristotelică, un cadru moral în care sunt rezolvate aspecte din vremea lui Aristotel pe care, desigur, le blamăm astăzi, precum statutul femeilor sau sclavia. Nu aș merge însă până la capăt cu zicala „virtuți noi pentru vremuri noi", mai ales dacă nu sunt ancorate în virtuțile clasice care și-au dovedit, în timp, relevanța. Adesea vedem astăzi confuzii între valori și virtuți, între idealuri și virtuți. Adevărul este o valoare, pe când onestitatea este o virtute (de caracter).

O abordare contemporană a eticii virtuții pe care o apreciez foarte mult este cea dezvoltată de filosoafa Shannon Vallor. Vallor propune o interpretare actuală a virtuților clasice, în contextul secolului XXI marcat de co-existența noastră tot mai strânsă și întrepătrunsă cu tehnologia, astfel că vorbește despre virtuți tehno-morale. Interpretarea lui Vallor are însă la bază o înțelegere profundă a virtuții aristotelice și nu se bazează pe un set de virtuți complet diferit de cel propus de Aristotel. Virtuțile tehno-morale sunt, pe scurt, acele virtuți care ne ajută să relaționăm înțelept cu tehnologia și să o folosim spre binele umanității. De exemplu, virtutea tehno-morală a curajului se referă la cultivarea unei dispoziții stabile de a aborda cu o doză adecvată de speranță și teamă pericolele și oportunitățile (morale și materiale) pe care le generează tehnologiile emergente.

4. Ați abordat problema responsabilității morale din perspectiva eticii virtuții în tradiție aristotelică. Asumând că alegerea între teoriile etice cunoscute nu a fost arbitrară, de ce considerați că etica aristotelică a virtuții este un cadru teoretic mai adecvat pentru înțelegerea noțiunii de responsabilitate morală decât, să spunem, etica deontologică sau cea utilitaristă?

Într-adevăr, am ales deliberat etica virtuții în tradiție aristotelică pentru a aborda tema responsabilității morale. Este un subiect care mă preocupă încă din perioada doctoratului, când am cercetat nivelul individual și nivelul colectiv de responsabilitate morală în context organizațional. Nu eram inițial foarte hotărâtă cu privire la cadrul etic pe care să îl folosesc, iar factorul determinant a fost cursul doctoral susținut de regretatul profesor Valentin Mureșan, coordonatorul meu de doctorat. A fost un curs dedicat *Eticii Nicomahice* și atunci am realizat că multe direcții contemporane de interpretare a conceptului de responsabilitate morală au de fapt rădăcini în discuția lui Aristotel despre virtuți și vicii și despre condițiile de posibilitate pentru acțiuni virtuoase sau vicioase. În funcție de îndeplinirea acestor condiții, avem temeiuri să blamăm sau să lăudăm o persoană pentru anumite acțiuni, altfel spus să o considerăm responsabilă moral. Aristotel nu folosește expresia „responsabilitate morală", însă este deja un

consens printre comentatorii săi că modul în care tratează atribuirea blamului sau laudei echivalează cu atribuirea responsabilității morale. Susan Sauvé Meyer are o carte excelentă pe acest subiect – *Aristotle on Moral Responsibility*, extrem de utilă pentru cei care sunt interesați să exploreze tema.

5. De regulă, în filozofia contemporană, etica virtuților este văzută ca o tradiție filozofică morală marginală în domeniul eticii. De ce credeți că celelalte două alternative, deontologia și utilitarismul, sunt mai răspândite – atunci când vine vorba despre cercetarea privind statutul moral al inteligenței artificiale în particular, dar și în general?

Așa este, etica virtuții este privită mai degrabă ca „a treia mare teorie etică normativă”, la distanță destul de mare de utilitarism și deontologismul kantian. Cred că unul dintre motivele care au condus la marginalizarea eticii virtuții a fost încărcătura scolastică pe care etica aristotelică a primit-o în perioada medievală. Conul de umbră în care s-a găsit apoi etica virtuții în secolele XVIII-XIX s-a extins până după jumătatea secolului XX, când revine în discuțiile academice după publicarea articolului celebru “Modern Moral Philosophy” semnat de Elizabeth Anscombe. Statutul marginal al eticii virtuții se schimbă tot mai mult spre începutul secolului XXI, iar acest lucru se vede inclusiv în zone ale eticii aplicate precum etica afacerilor și etica inteligenței artificiale, unde tot mai mulți cercetători văd avantajele eticii virtuții pentru a analiza, printre altele, statutul moral al IA sau modul în care contextul organizațional influențează comportamentul moral al membrilor organizației.

6. Există numeroase argumente care susțin posibilitatea – și chiar inevitabilitatea – dezvoltării sistemelor de inteligență artificială generală sau *strong AI*, care se disting de sistemele de inteligență artificială cu sarcini particulare/specifice sau *narrow AI*. Totodată, există la fel de multe argumente care resping această posibilitate – sau măcar inevitabilitatea. Având în vedere mai ales literatura de specialitate pe care vă bazați în proiectul de cercetare menționat mai devreme, ce ne desparte astăzi de posibilitatea de a crea un sistem de inteligență artificială generală?

Eu mă regăsesc în tabăra celor care privesc cu mult scepticism posibilitatea de a vorbi de o inteligență artificială generală și cred că ar trebui să fim mai degrabă atenți la modul în care oamenii pot folosi IA pentru a face rău – intenționat sau nu -, pentru că suntem deja într-un punct în care pot fi generate efecte negative pe scară largă prin intermediul IA. Îmi e greu să spun cu certitudine dacă este complet imposibil, dar cu siguranță inteligența artificială generală nu este ceva inevitabil. Această imagine a unei

IA perfect echivalente sau chiar superioare ființei umane în toate aspectele (nu doar în anumite zone specifice, cum este cazul unor sisteme de IA dezvoltate deja în prezent, precum AlphaZero) a fost promovată intens în media, fie prin intermediul filmelor, fie chiar și prin discursuri ale unor persoane publice active în zona de tehnologie sau chiar ale unor cercetători, care au supralicitat diverse contexte. Totuși, astăzi sunt destul de puțini cercetători serioși care își dedică timpul și resursele pentru a dezvolta o inteligență artificială generală.

Sunt multe lucruri care ne despart de AGI (*Artificial General Intelligence*). Cred că este o linie de demarcație tare, o limită pe care IA nu o poate depăși și care include atât cerințe ontologice, cât și epistemologice, de la conștiință, auto-reflexivitate și capacitatea de a simți, până la intenționalitate, urmărirea unui scop auto-impus și coordonare complexă în interacțiunea cu mediul – căci AGI ar trebui să aibă și o formă fizică, o structură materială corporală, chiar dacă una foarte diferită de cea biologică. Pentru toate acestea cred că ar fi necesar un salt non-linear calitativ, și nu doar unul cantitativ, față de sistemele artificiale autonome pe care le avem în prezent.

7. Traversăm niște vremuri foarte interesante pentru domeniul eticii aplicate, dacă ne uităm, de pildă, la problemele generate de pandemie, la război sau, mai recent, la decizia Curții Supreme din Statele Unite ale Americii privind avortul. Cum poate învăța un tânăr, care a deschis ochii în lumea Internetului și a rețelelor de socializare, să mai distingă astăzi între virtuțile morale clasice și somarea de a-și *semnaliza virtuțile* de tot soiul la care este supus?

Este o ironie de situație faptul că *semnalizarea* propriilor virtuți sau îndemnul la *semnalizarea* virtuților celorlalți reprezintă în sine un act lipsit de virtute. Cred că un cuvânt cheie în demersul de a distinge între virtuți și pseudo-virtuți este efortul. Tinerii au nevoie să revalorizeze efortul, munca, stăruința. Virtuțile, mai cu seamă cele de caracter, se dobândesc cu mult efort, adesea trecând prin întâmplări mai puțin fericite. Dacă vrei să fii cu adevărat o persoană mai bună, cu un sentiment al împlinirii, sentiment ce cuprinde modul în care relaționezi cu sine și cu cei din jur, atunci trebuie să accepți că ai mult de muncit, că acest lucru nu se va întâmpla peste noapte, că este un drum cu mărăcini și jelituri, dar și cu cer senin și bucurie. Rețelele virtuale de socializare pot contribui la îmbogățirea rețelei noastre sociale, dar nu o pot substitui, iar acest efect de susținere se obține numai în măsura în care continuăm relaționarea cu ceilalți în lumea reală. Prieteniile se construiesc în timp și cu efort, cer dedicare și implicare prin fapte concrete, iar uneori asta înseamnă să te trezești în miez de noapte să îți iei un prieten de la aeroport și alteori să împarți ultima felie de pâine în vârf de munte. De aceea, prietenia deopotrivă este o virtute și necesită virtute, cum spune Aristotel. Și de aceea prietenia adevărată, la fel ca virtuțile adevărate, le aduc cu

adevărat împlinire celor care depun efortul necesar pentru a le pune în practică.

Imagine: Arhiva personală