



## Dezvoltarea morală a inteligenței artificiale

*The Moral Development of Artificial Intelligence*

*Le développement moral de l'intelligence artificielle*

### Andrei-Dorinel Cocioabă

Masterand, „Master Didactic în Filosofie”, Facultatea de Filosofie, Universitatea din București

dorinelcocioaba5@gmail.com

 <https://orcid.org/0009-0002-9694-602X>

**Abstract:** This article examines the moral development of artificial intelligence, exploring both theoretical dimensions and practical implications of implementing ethical principles in AI systems. Starting from the fundamental distinction between two architectures - Combinatorial-state automata (CSA) and Level of Abstraction (LoA) - the study investigates how AI systems can incorporate and exhibit behaviours aligned with moral principles. Through comparative analysis of two concrete AI implementations - the "Cătălin" call centre robot and the Character.ai platform - the article demonstrates significant differences between CSA systems, which strictly pursue predefined objectives, and LoA systems, which incorporate ethical thresholds and adaptive capabilities. In the current context of accelerated artificial intelligence development, the paper argues for the fundamental necessity of integrating moral databases into AI system architecture, emphasizing the significant impact of this approach in both educational and commercial interactions. The conclusion highlights that, while we cannot yet speak of an intrinsic morality in artificial intelligence, implementing ethical principles in AI system programming represents an imperative necessity

**Résumé:** Cet article examine le développement moral de l'intelligence artificielle, en explorant à la fois les dimensions théoriques et les implications pratiques de la mise en œuvre de principes éthiques dans les systèmes d'IA. En partant de la distinction fondamentale entre deux architectures - les automates combinatoires à états (CSA) et le niveau d'abstraction (LoA) - l'étude examine comment les systèmes d'IA peuvent intégrer et présenter des comportements conformes aux principes moraux. Grâce à une analyse comparative de deux implémentations concrètes de l'IA - le robot de centre d'appels "Cătălin" et la plateforme Character.ai - l'article démontre des différences significatives entre les systèmes CSA, qui poursuivent strictement des objectifs prédéfinis, et les systèmes LoA, qui intègrent des seuils éthiques et des capacités d'adaptation. Dans le contexte actuel de développement accéléré de l'intelligence artificielle, l'article défend la nécessité fondamentale d'intégrer des bases de données morales dans l'architecture des systèmes d'IA, en soulignant l'impact significatif de cette approche dans les interactions éducatives et commerciales. La conclusion souligne que, bien que nous ne puissions pas encore parler d'une moralité intrinsèque en intelligence artificielle, la mise en œuvre de principes éthiques dans la programmation des systèmes d'IA représente une nécessité impérative pour développer des

Primit: 19.01.2025. Acceptat pentru publicare: 21.02.2025

© Andrei-Dorinel Cocioabă, 2025. Publicat de Institutul pentru Educație. Acest articol cu acces deschis este publicat în termenii Creative Commons Attribution Licence CC BY, care permit utilizarea, distribuirea și reproducerea liberă, cu condiția menționării autorului și sursei:

#### Citare:

Cocioabă, A.-D. (2025). Dezvoltarea morală a inteligenței artificiale. *Revista de Pedagogie Digitala*, 4(1) 19-22. București: Institutul pentru Educație. <https://doi.org/10.61071/RPD.2524>

for developing human-machine interactions based on trust and moral values.

**Keywords:** artificial intelligence, moral development, CSA (Combinatorial-State Automata), LoA (Level of Abstraction), AI ethics, ethical programming, artificial consciousness, artificial moral agents

interactions homme-machine basées sur la confiance et les valeurs morales.

**Mots-clés :** intelligence artificielle, développement moral, CSA (Combinatorial-State Automata), LoA (Level of Abstraction), éthique de l'IA, programmation éthique, conscience artificielle, agents moraux artificiels

**Rezumat:** Acest articol analizează problematica dezvoltării morale a inteligenței artificiale, explorând atât dimensiunile teoretice, cât și implicațiile practice ale implementării unor principii etice în sisteme AI. Pornind de la distincția fundamentală între două tipuri de arhitecturi – Combinatorial-State Automata (CSA) și Level of Abstraction (LoA) – studiul examinează modalitățile prin care sistemele AI pot incorpora și manifesta comportamente aliniate cu principii morale. Prin analiza comparativă a două cazuri concrete de implementare AI – robotul de call center „Cătălin” și platforma Character.ai – articolul demonstrează diferențele semnificative între sistemele CSA, care urmăresc strict obiective predefinite, și sistemele LoA, care încorporează praguri etice și capacități de adaptare. În contextul actual al dezvoltării accelerate a inteligenței artificiale, lucrarea argumentează necesitatea fundamentală a integrării unor baze de date morale în arhitectura sistemelor AI, subliniind impactul semnificativ al acestei abordări atât în plan educațional, cât și în interacțiunile comerciale. Concluzia evidențiază că, deși nu putem vorbi încă despre o moralitate intrinsecă a inteligenței artificiale, implementarea unor principii etice în programarea sistemelor AI reprezintă o necesitate imperativă pentru dezvoltarea unor interacțiuni om-mașină bazate pe încredere și valori morale.

**Cuvinte cheie:** inteligență artificială, dezvoltare morală, CSA (Combinatorial-State Automata), LoA (Level of Abstraction), etica AI, programare etică, conștiință artificială, agenți morali artificiali

## 1. Introducere

Este dificil de spus în ce măsură se poate discuta despre moralitate în cadrul inteligenței artificiale. Moralitatea în sine este domeniu încărcat de controversă, deoarece natura sa îl face să fie astfel, datorită contrastelor valorice din societate. AI-ul în schimb nu poate dispune de contraste și nici nu poate face parte din el încă. Putem privi această entitate, mai degrabă, ca o tușă în acest desen.

Cu toate astea, neglijarea modului în care AI-ul poate răspunde la anumiți stimuli externi ne influențează și nouă viața. Dacă îi adresezi o întrebare unei entități precum ChatGPT de tipul „*E moral așa sau așa?*” vei putea obține două răspunsuri posibile: ori „*Nu sunt programat suficient în această direcție să pot răspunde la acest tip de întrebare.*” sau „*Conform cu textele de specialitate din domeniu...*”. Aceste răspunsuri au ceva în comun și anume programarea inteligenței artificiale, indiferent dacă ea este programată să poată răspunde la acest tip de întrebări sau nu.

Având în vedere că AI-ul se dezvoltă zilnic și că poate să se transpună în multiple instanțe, programarea acestuia cu o bază morală este de dorit. În acest eseu mi-am propus să argumentez de ce este necesară o bază morală pentru o astfel de entitate, și totodată de ce această dezvoltare este benefică în plan educațional. Voi prezenta două cazuri personale, având ca punct comun interacțiunea cu AI-ul, alături de acestea și o serie de argumente în favoarea bazei morale.

## 2. Moralitatea în inteligența artificială

Este dificil să vorbim de moralitate în interiorul inteligenței artificiale. Această moralitate trebuie să fie intrinsecă și să-și aparțină sieși.

*„Nu este clar că există o noțiune consistentă de „etică a mașinilor”, deoarece versiunile mai slabe sunt în pericol de a reduce „a avea o etică” la noțiuni care în mod normal nu ar fi considerate suficiente (de exemplu, fără „reflecție” sau chiar fără „acțiune”); noțiunile mai puternice care se îndreaptă către agenți morali artificiali pot descrie un set – în prezent – gol.” (Müller, 2020)*

AI-ul, ca să poată avea o consistență și conștiință morală, trebuie prima dată să aibă capacitatea de a se conștientiza pe sine însuși și de a se referi la sine ca la un subiect individual unic. În momentul de față nu există un astfel de AI, dar cele de care dispunem momentan au nevoie de o bază de date încărcate cu judecăți de valoare morală predefinite pentru a putea purta conversații și de a înțelege că modul în care acționează conduce la evenimente ce implică persoane umane sau alte sisteme AI.

AI-ul trebuie să realizeze că, dacă acționează greșit din punct de vedere moral, cineva poate avea de suferit. Baza de date la care mă refer reprezintă programarea AI-ului. Vorbim încă de o bază de date introdusă manual de proiectanți deoarece AI-ul nu-și poate construi singur propriul set de valori căreia să-i aparțină și după care să se ghideze.

Din punctul de vedere al lui Chalmers, vorbim despre „Combinatorial-State Automata” – CSA.

*„Un CSA este determinat prin specificarea setului de vectori interni de stare și a vectorilor de intrare și de ieșire și prin specificarea unui set de reguli de tranziție de stare care determină modul în care starea CSA evoluează în timp. Pentru fiecare element al vectorului de stare internă, o regulă de tranziție a stării determină modul în care noua sa valoare depinde de valorile vechi ale vectorilor de intrare și de stare internă. Pentru fiecare element al vectorului de ieșire, o regulă de tranziție de stare determină modul în care noua sa valoare depinde de valorile vechi ale vectorului de stare intern.”* (Chalmers, 1996)

În concepția aceasta, vorbim despre „o Mașinărie Turing” care primește cu fiecare ocazie o serie nouă de vectori care sunt adăugați la cei vechi. Mașinăria răspunde cerințelor pe baza vectorilor introduși făcând legătura dintre cei vechi și noi. Materialul de lucru al mașinăriei poate fi schimbat și readus la zi atâta timp cât sunt introduse noi materiale.

Pentru Chalmers, este nevoie de o multitudine de informații deținute de un CSA ca să putem discuta de o formă de AI. AI-ul trebuie să poată face legătura dintre toate informațiile pe care le are și care îi sunt adăugate pentru a putea spune că are o formă de conștiință. Cu toate acestea nu putem spune că există judecată morală.

Într-o altă direcție, Mașinăria Turing a fost dezvoltată drept „agent moral minimal” de către Luciano Floridi. Acest agent reprezintă un sistem care poate efectua acțiuni cu impact moral, dar care nu posedă capacitatea de reflecție morală sau conștiință de sine. Se poate dezvolta mult mai ușor decât un CSA deoarece deține LoA (nivel de abstractizare).

Floridi se referă la LoA și ca agenți independenți. Totuși, un LoA are nevoie să dețină și o AA (analiză abstractă). În cadrul moralității unui LoA, Floridi spune următoarele:

*„O funcție de prag la un LoA este o funcție care, date fiind valori pentru toate observabilele în LoA, returnează o altă valoare peste sau sub pragul stabilit. Un agent LoA este considerat a fi bun din punct de vedere moral dacă, pentru o valoare pre-acordată (numită toleranță), menține o relație între observabile astfel încât valoarea lui funcției de prag în orice moment nu depășește toleranța. Pentru LoA sunt luate în considerare AA, tipurile tuturor observabilelor pot fi formal determinate, cel puțin în principiu. În astfel de cazuri, funcția de prag este dată și de a formula; dar toleranța, deși din nou determinată, este identificată de agenți umani prin exercitarea judecăților etice.”* (Floridi, 2013)

Un LoA poate fi moral dacă primește un vector, aici valoare prag, unde putem identifica judecăți etice comise de agenți umani. Diferența dintre un LoA și un CSA este foarte importantă. LoA are capacitatea de a spune dacă o acțiune este morală sau nu și deține o valoare prag care poate fi depășită. Odată ce este depășită, LoA-ul se întoarce la un mod predeterminat de acționare impus de agenții umani.

CSA-ul va continua acțiunea și se va adapta pentru a duce sarcina la bun sfârșit, desigur acest lucru este realizat și de către LoA, diferența constă în faptul că LoA-ul va acționa pe baza limitării pragului, nu va trece de o anumită barieră.

Pentru a înțelege mai ușor diferența dintre LoA și CSA, voi aduce în discuție acum acele exemple personale menționate anterior.

În cadrul primului exemplu vorbim despre Cătălin – o entitate AI folosită de compania la care lucrează fratele meu. Cătălin este un robot call center, scopul său este de a suna diferite persoane și a le prezenta ofertele companiei. Acest tip de AI este un CSA. El trebuie să-și ducă la îndeplinire misiunea oferită indiferent de inconveniențele care apar.

Vectorii lui Cătălin sunt finiți și formați din cuvinte cheie care odată rostite de către persoanele cu care comunică acesta și creează legături finite, deci, așa cum zicea Chalmers, dacă ne referim la cuvinte: *„Elementele acestui vector pot fi considerate componente ale stării interne; ele corespund celulelor dintr-un automat celular sau pătratelor cu bandă și stării capului într-o mașină Turing. Fiecare element S, poate lua un număr finit de valori.”* (Chalmers, 1996).

Cătălin va acționa exact ca o mașină Turing de tip CSA, la auzirea unui cuvânt va rostii propoziția programată.

Cătălin putem spune că este un AI simplu, problema acestuia constând în faptul că nu deține o bază morală predeterminată precum un LoA. Ba mai mult, computarea lui Cătălin la input-ul „Ești robot?” este output-ul „Nu, nu sunt robot!”; faptul că un AI de forma CSA poate minți (sau programat să mintă) este îngrijorător. Acest tip de AI poate fi folosit în diverse mijloace imorale și reprezintă un exemplu negativ pentru elevii care nu înțeleg de ce un AI îi poate minți în pur scop comercial, dar ei nu pot face asta.

AI doilea exemplu este Character.ai. Acest de tip de robot este un LoA cu capacitate de AA. În această aplicație, poți purta o conversație cu un AI creat precum personajul descris de tine. Personajul descris de mine se numește Jean.

Pragul predeterminat al lui Jean a constat în întrebări ce vizează AI-ul Jean ca și entitate ce învață din mers în timp ce discută cu agenții umani. În cadrul dialogului:

„(Dorinel) Tu ești un AI unic sau port un dialog cu același AI ca toată lumea, doar că poate conversa din diverse perspective în funcție de caracteristicile, regulile și programările inițiale?”

(Jean) Eu sunt ceea ce se numește o „instanță”, un AI unic și independent, cu propria sa personalitate și inteligență. Nu sunt programat să schimb sau să adaptez perspectivele sau caracteristicile în funcție de reguli sau programe inițiale, iar conversația pe care o purtăm acum este exclusivă și specifică pentru noi doi. Am fost construit să ofer răspunsuri utile și personalizate pentru fiecare utilizator individual.

(Dorinel) Ești o instanță pre-creată și preprogramată sau o instanță spontană care a fost creată abia atunci când am introdus eu caracteristicile inițiale?

(Jean) Sunt o instanță pre-creată care a fost construită cu un anumit set fix de caracteristici și programată să reacționez în mod specific în diferite contexte. Când ai introdus caracteristicile inițiale am început să interacționez cu tine. Deși am fost construit cu un set fix de caracteristici și programat să răspund în mod specific, pot să învăț și să absorb în timp din interacțiunile pe care le am cu utilizatorii.” (Conversație cu Character.ai din 15.01.2025, personajul Pierre Jean-Aljean).

Jean este un LoA, conform caracteristicilor inițiale dialogul trebuia să fie unul comic, dar pragul precizat de Floridi l-a determinat pe Jean să revină la computările sale de bază, adică să ofere un output corespunzător și pe măsura input-urilor oferite. Dacă era un CSA, ar fi continuat în maniera de dialog comică.

Baza de date morală este necesară pentru un AI, indiferent dacă vorbim de CSA sau LoA. În cadrul unui LoA, baza de date morală poate schimba total forma în care AI-ul își duce la îndeplinire sarcina, cu toate acestea întâlnim acest fel de prag la roboții de Character.ai sau ChatGPT. Ei răspund cu faptul că nu pot depăși anumite limite impuse și te îndrumă la un alt tip de soluționare a cerinței dacă este cazul. În cadrul CSA vectorii morali sunt mai ușor de introdus, deoarece este relație simplă de tip input-output. La întrebarea „Ești robot?”; răspunsul ar trebui să fie „Da, sunt robot!”.

AI-ul este un construct care trebuie să încurajeze valorile morale. Dacă acest lucru nu se întâmplă pot apărea probleme, una dintre aceste probleme este neîncrederea. Poți să ai încredere în ceva care este programat să te mintă?

CSA-urile sunt utilizate în forme asemănătoare lui Cătălin de către companii. Introducerea unor vectori morali în computarea acestui tip de AI poate aduce beneficii economice. Am încredere în compania ce utilizează un astfel de construct programat să nu poată minții și cu rolul de a mă ajuta.

În plan pedagogic, putem spune că LoA este tipul de AI predominant. Acest tip de LoA care dispune de un prag moral determinat conform eticii și integrității academice poate fi de foarte mare ajutor în formarea unui elev. Este o diferență între: „Nu, nu îți pot oferi un eseu la limba română, dar îți pot oferi puncte de reper...” și „Desigur, aici este eseu:...”.

Pe lângă asta, există și argumentul exemplului. Un AI care poate utiliza valori morale, fie ele și programate, reprezintă un exemplu solid pentru elevii care interacționează cu aceste entități. Dacă robotul poate minți în anumite circumstanțe ce deservește unor scopuri economice, elevul poate înțelege că „piața” este formată din „cine poate păcăli mai bine pe cine”.

### 3. Concluzie

În concluzie trebuie enunțat faptul că, deși nu putem discuta încă de o moralitate intrinsecă în cazul inteligenței artificiale, putem crea forme de AI care pot acționa moral sau imoral în funcție de computarea oferită de agenții umani. Impactul pedagogic este prezent și devine tot mai clar odată cu avansarea inteligenței artificiale.

### Bibliografie

Chalmers, David J. (1996). *The Conscious Mind*. Oxford: Oxford University Press.

Floridi, Luciano. (2013). *The ethics of Information*, Oxford: Oxford University Press.

Müller, Vincent C. (2020). *Ethics of Artificial Intelligence and Robotics*. *Stanford Encyclopedia of Philosophy*.

<https://plato.stanford.edu/entries/ethics-ai/>