UNIVERSITATEA „ALEXANDRU IOAN CUZA” DIN IAȘI
FACULTATEA DE FILOSOFIE ȘI ȘTIINȚE SOCIAL-POLITICE
ȘCOALA DOCTORALĂ
Domeniul: Filosofie

INTELIGENȚA ARTIFICIALĂ ȘI PROBLEMA CONȘTIINȚEI

REZUMATUL TEZEI DE DOCTORAT

Conducători de doctorat:

PROF. UNIV. DR. GEORGE BONDOR

PROF. UNIV. DR. IOAN ALEXANDRU TOFAN

Student-doctorand:

ALEXANDRA CHIRILĂ

Iași

2026

CUPRINS

În 1997, DeepBlue l-a învins pe Gari Kasparov la șah, iar lumea a asistat la prima instanță în care un computer a câștigat un joc împotriva unui campion mondial. De atunci și cu precădere în ultimul deceniu, inteligența artificială a trecut de la un domeniu cu impact limitat la sisteme care acaparează discursul public și care par deosebit de inteligente, aproape conștiente. Dezvoltarea modelelor de învățare automată (*machine learning*) și, mai recent, a rețelelor neuronale de tip deep learning, cât și a arhitecturii transformer a condus la performanțe uluitoare în recunoașterea de tipare și generare de text. Acest progres a alimentat atât speranțe în ce privește posibilitatea construcției unui omolog artificial, dar și temeri cu privire la impactul social și etic al unor astfel de sisteme.

Astfel, cercetarea de față pornește de la întrebarea *poate inteligența artificială să fie conștientă și intențională?* Această întrebare se află la intersecția dintre filosofie, neuroștiințe și tehnologie și a devenit populară odată cu dezvoltările din domeniul învățării automate. Tema este relevantă atât pentru

filosofia minții, dar și pentru cercetările din domeniul AI (*Artificial Intelligence*), deoarece plasează în centru întrebarea dacă simularea cognitivă ar putea să fie suficientă pentru conștiință sau este nevoie de aspecte precum corporalitate și stări afective.

Stadiul cercetărilor în domeniu

După cum am menționat, inteligența artificială a avut parte de o dezvoltare semnificativă pe parcursul ultimilor ani, trecând de la arhitectura simbolică la modele predictive bazate pe rețele neuronale artificiale, capabile de o multitudine de sarcini, cum ar fi generarea de text sau imagini. Cercetările incipiente s-au concentrat pe abordări simbolice, modele numite și *inteligență artificială de modă veche* sau GOFAI (*good old fashioned AI)*. Programe precum Cyc reprezintă încercarea de a concentra vaste cunoștințe umane, însă acestea erau limitate la manipulare simbolică conform regulilor logice, dar nu aveau capacitatea de a înțelege datele cu care operau. Apoi, modelele bazate pe rețele neuronale artificiale au făcut posibilă recunoașterea de tipare în date. Modelele

contemporane de inteligență artificială, cum ar fi *învățarea profundă* (*deep learning*) și arhitectura transformer au îmbunătățit considerabil performanțele programelor bazate pe rețele neuronale artificiale. Modelele lingvistice mari (*large language model* - LLM), cum ar fi ChatGPT sau Gemini pot genera texte coerente și pot simula dialogul, iar sistemele de învățare prin recompensă (*reinforcement learning*) pot învinge campioni mondiali la jocuri precum go sau Dota 2. Aceste sisteme funcționează prin intermediul mecanismelor predictive și generează outputuri (răspunsuri) în funcție de tipare identificate în datele pe baza cărora au fost antrenate, însă performanțele lor nu se datorează înțelegerii semantice a acestor date, după cum va fi discutat.

Pentru că întrebarea de cercetare implică și conștiința, trebuie menționat că acest fenomen va fi înțeles drept experiență trăită, însoțită de stări afective cu miză care ghidează comportamentul și luarea deciziilor. Pentru oameni, această trăsătură se află în strânsă legătură cu cogniția. Cogniția implică percepție, atenție, memorie, iar acestea sunt modulate

de stările afective și de nevoi ale corpului. Studiile recente din sfera neuroștiinței sugerează că afectele reprezintă elemente constitutive ale conștiinței, nu fenomene secundare, așa cum s-a crezut o bună parte din istoria acestui domeniu. Întrebarea vizează și intenționalitatea, în calitate de orientare către lume cu un scop, iar această trăsătură este inseparabilă de experiențele conștiente și afective. Spre exemplu, curiozitatea îndeamnă la explorare, iar frica la evitare, dar lipsa acestor stări face ca acțiunile în lume să nu poată fi orientate. Inteligența artificială, în forma ei actuală, nu posedă nici conștiință, nici intenționalitate. Ca atare, deși pot simula înțelegere sau intenționalitate, sistemele artificiale nu pot înțelege semantic informațiile procesate și nici nu se pot orienta către lume. Prin plasarea cercetărilor din domeniul AI în sfera conștiinței și a afectelor, lucrarea de față își propune să argumenteze că modelele artificiale pot funcționa conform așteptărilor, însă nu posedă o dimensiune fenomenologică, semantică sau intențională.

Inteligența artificială și fenomenul conștiinței au fost, pentru o perioadă lungă de timp, tratate separat. Primul domeniu a fost cercetat cu precădere în contexte tehnice ce vizează să îmbunătățească capacitățile acestor sisteme, cu scopul de a concretiza un proiect ce poartă numele de AGI sau inteligență artificială generală. Acest proiect presupune construcția unui model capabil de aceleași procese cognitive precum oamenii, printre care amintim adaptabilitatea, generalizarea sau abstractizarea. În ultimii ani s-a observat o încetinire a procesului de scalare, adică a sporirii capacităților computaționale, fapt ce denotă necesitatea unei schimbări în ce privește modul în care construim astfel de sisteme. Pe de altă parte, cercetările cu privire la conștiință s-au concentrat în mare parte pe abordările funcționaliste. Pentru funcționalism, stările mentale sunt definite de ceea ce fac, nu de substratul biologic care le cauzează. Așadar, pentru susținătorii funcționalismului contează funcția stărilor mentale, nu din ce sunt alcătuite aceste stări. Problema abordării funcționaliste e că plasează în plan secundar dimensiunea subiectivă a stărilor

mentale, astfel, funcționalismul apare ca o altă formă a behaviorismului. Mai mult decât atât, cercetările ce vizează fenomenul conștiinței au subestimat importanța stărilor afective, însă studiile recente au revenit asupra acestei ipoteze și au contribuit la noțiunea că afectele reprezintă elemente fundamentale ale conștiinței.

Astfel, dacă fenomenul conștiinței este unul ce comportă o puternică dimensiune afectivă, iar proiectul AGI se revendică din capacitățile cognitive umane, atunci cercetările în acest domeniu trebuie să implice posibilitatea implementării stărilor afective în cadrul sistemelor artificiale.

Lacunele cercetărilor contemporane

În ciuda complexității răspunsurilor generate de sistemele artificiale contemporane, există o lacună semnificativă, anume faptul că aceste sisteme nu au acces la dimensiunea semantică și intențională a lumii. Majoritatea cercetărilor din acest domeniu prioritizează optimizarea răspunsurilor sau capacitatea predictivă a AI-ului în detrimentul dezvoltării

inteligenței reale. Outputurile sistemelor artificiale de astăzi pot induce utilizatorul în eroare, deoarece aceste sisteme mimează cu succes abilități specific umane, cum ar fi dialogul, oferirea răspunsurilor coerente și corecte din punct de vedere sintactic etc. Astfel, se încearcă simularea inteligenței sau a stărilor afective fără ca acestea să fie însoțite de mize reale. Spre exemplu, inferențele bayesiene sunt folosite pentru a formaliza raționamentele abductive, deși acestea presupun un proces inerent creativ, acela de a formula cea mai bună explicație. Metoda *reinforcement learning* este privită ca potențială abordare pentru a simula stări afective cu valență, însă în lipsa dimensiunii semantice, valorile numerice (care au rolul de a evalua și ghida comportamentele sistemelor artificiale) nu înseamnă nimic pentru AI, iar creativitatea necesară pentru efectuarea raționamentelor abductive nu poate apărea în lipsa unui model intern al lumii și fără experiența trăită a acestui model. În ce privește conștiința, aceasta este privită predominant în calitate de eveniment exclusiv cerebral, iar dimensiunea corporală este ignorată.

Afectele ce contribuie la acest fenomen implică stări resimțite în corp și stări homeostatice care monitorizează funcționarea internă a organismului.

În cercetările contemporane, procesarea predictivă în cadrul inteligenței artificiale este privită în cheie behavioristă. Mai exact, LLM-urile reprezintă programe software cu funcționalități în sfera limbajului, folosind tehnologii de tip *machine learning*. Acestea operează cu unități de date ce poartă numele de *tokenuri*, care reprezintă caractere dintr-un text, fie ele cifre sau litere, iar în stadiul actual, LLM-urile sunt limitate la predicții în legătură cu ce caracter este cel mai probabil să urmeze într-un șir de token-uri. Pentru că sintaxa poate fi transpusă în limbaj computațional, aceste programe pot genera cu succes fraze și texte care au sens gramatical și logic, însă semantica nu poate fi transpusă în limbaj computațional, ceea ce înseamnă că aceste modele de limbaj nu au capacitatea de a înțelege inputul primit sau outputul generat, întocmai precum un model sofisticat al Camerei Chinezești. Aspectul behaviorist al dezvoltării sistemelor artificiale constă în faptul că

evaluarea calității outputurilor nu reprezintă evaluarea competențelor reale pe care le posedă aceste sisteme. Concret, procesarea predictivă în cadrul AI este privită cu rolul de a prezice următorul *token* în serie, pe când acest tip de procesare are rolul de a minimiza nivelul de energie liberă din organism, în cazul oamenilor. Procesarea predictivă de care este capabilă inteligența artificială nu are rolul de a consolida sau modifica modele descriptive interne ale lumii. Mai mult decât atât, acest tip de procesare este efectuată proactiv și intențional în cazul oamenilor, pe când sistemele artificiale sunt pasive în procesul de prezicere a următorului *token* în serie. AI nu poate procesa informații decât atunci când i se indică să facă asta. Pentru oameni, procesarea predictivă mai are rolul de a ghida comportamentul, de a anticipa nevoi survenite prin intermediul stărilor homeostatice și, mai ales, contribuie activ la ce am putea numi trăsături superioare ale cogniției, cum ar fi imaginația și planificarea.

Totuși, în continuare există perspective ce consideră inteligența artificială bazată pe date un pas

în față pentru dezvoltarea inteligenței artificiale generale, precum cea a lui Stuart Russell, care discută despre problema alinierii scopurilor inteligenței artificiale cu ale omenirii, problemă care amintește de conceptul convergenței instrumentale propus de Nick Bostrom. Însă, când vine vorba despre etică în construcția inteligenței artificiale, discursul se concentrează în principal pe limitarea biasurilor sau a dezinformării răspândite cu ajutorul acestor sisteme și mai puțin pe implicațiile antropomorfismului. Această tendință presupune extinderea intenționalității și a capacităților cognitive proprii și proiecția acestora în artefactele din jurul nostru, însă poate avea consecințe negative asupra modului în care interacționăm cu programele de AI, investindu-le cu expertiză pe care nu o au.

Contribuțiile lucrării de față

Pe parcursul cercetării vom argumenta că stările afective sunt elemente constitutive a fenomenului conștiinței și, mai mult, că funcționalismul este necesar, dar nu și suficient pentru a explica acest fenomen. Este un demers util să

analizăm funcțiile structurilor cerebrale, dar ele nu pot explica stările experiențiale, înțelegerea sau conștiința ca întreg. Funcționalismul nu poate reprezenta o teorie completă a conștiinței în lipsa aspectelor precum întruparea sau stările afective. Vom argumenta că și semnificațiile se află în strânsă legătură cu afectele. Înțelegerea lumii apare datorită existenței nevoilor biologice, a vulnerabilității (în sens de corp care trece prin procese inevitabile de degradare) și a mizelor legate de supraviețuire. Mai mult, dacă funcționalismul este privit drept teoria corectă a mecanismelor conștiinței, influența sa depășește domeniu filosofiei și determină modalitățile prin care construim și evaluăm sistemele artificiale. Pentru susținătorii acestei perspective, ceea ce contează este rolul pe care un sistem îl îndeplinește și nu existența experiențelor subiective. Acest lucru contribuie la ideea că reproducerea aceleași funcții este suficientă pentru a atribui inteligență sau conștiință unui sistem artificial. Totuși, dacă inteligența umană presupune intenționalitate și experiențe trăite, atunci abordarea strict funcționalistă ne poate induce în eroare,

deoarece riscă să reducă inteligența la o imitație a comportamentelor. După cum nota Alan Turing, o mașină computațională nu trebuie să *știe* ce este aritmetica, însă o inteligență artificială generală trebuie să știe ce este aritmetica pentru a o înțelege.

În ce privește contribuțiile în domeniul inteligenței artificiale, vom face o distincție între conștiință și intenționalitate când vine vorba de astfel de sisteme. Această distincție este necesară pentru că, în teorie, emergența unei proto-intenționalități artificiale este posibilă. Prin acest termen ne referim la capacitatea AI-ului de automonitorizare și de a întreprinde acțiuni cu rolul de a îndeplini nevoi legate de consum de energie, spre exemplu. Această proto-intenționalitate nu garantează apariția semnificațiilor, iar conștiința implică accesul la dimensiunea semantică a lumii. De aceea, un sistem artificial (proto-)intențional nu este și conștient și, prin urmare, nici inteligent în sensul uman al termenului. Pe parcursul cercetării vom efectua o analiză critică a abordărilor care încearcă să simuleze întruparea în cadrul AI-ului. Vom argumenta că nevoile simulate nu

au consecințe reale și, deci, nici mize. Ca atare, aceste simulări nu pot produce semnificații la care sistemele artificiale să aibă acces. Înțelegerea este inseparabilă de mize resimțite ca stări valențiate, iar orice nevoie simulată nu va avea efecte atât timp cât inteligenței artificiale *nu îi pasă* de nevoile care îi sunt implementate. De asemenea, vom propune un set de condiții minim necesare pentru dezvoltarea proto-intenționalității în AI. Modelul nu are pretenția de a fi unul complet și este unul pur conceptual, dar poate servi drept punct de plecare pentru cercetarea intenționalității artificiale.

Nu în ultimul rând, vom avea în vedere o critică filosofică a discursului legat de singularitate, adică fenomenul ce presupune trecerea cauzală într-un ritm accelerat de la AGI la superinteligență (ASI). Vom argumenta că singularitatea, împreună cu celelalte scenarii speculative se revendică din tendința de antropomorfizare și nu din realitatea evoluțiilor tehnologice din ultimii ani. Pe parcursul lucrării, vom sublinia faptul că inteligența și înțelegerea (semantică) pot apărea doar atunci când există intenționalitate, iar

intenționalitatea apare doar atunci când există stări afective derivate din mecanismele homeostatice. Deviațiile homeostatice au miză, ele denotă un risc la adresa bunăstării, iar unele stări afective pun în pericol chiar viața organismului. Ca atare, fără un corp muritor, nu există afecte cu valență, iar afectele generează semnificații, chiar dincolo de stările homeostatice. Fără corp nu există deviație, fără deviație nu există afecte, iar fără afecte nu apare nevoia de supraviețuire. Fără aceste nevoi, nu există semnificație și fără semnificație nu există înțelegere. Iar fără înțelegere nu putem vorbi despre inteligență, deci nici despre AGI.

Importanța cercetării

Importanța acestei cercetări poate fi derivată din întrebarea centrală: *poate inteligența artificială să fie conștientă și intențională?* Întrebarea devine din ce în ce mai relevantă în condițiile popularizării și a progresului rapid al sistemelor artificiale bazate pe învățare automată.

Din perspectivă filosofică, cercetarea aduce contribuții la dezbaterea privind natura conștiinței și a intenționalității. Teoriile funcționaliste și behavioriste au subestimat dimensiunea subiectivă a stărilor mentale și au privit afectele în calitate de fenomene secundare. Pe de altă parte, studiile recente din domeniul neuroștiințelor arată că stările afective poartă un rol deosebit de important pentru conștiință și influențează trăsături precum cogniția și luarea deciziilor. Prin analiza legăturii dintre semnificație și afecte, lucrarea de față propune o perspectivă critică asupra funcționalismului și subliniază importanța experienței subiective și a corpului în accesul la semnificații.

Din perspectivă tehnologică, lucrarea de față este relevantă deoarece sistemele de inteligență artificială nu au acces la dimensiunea semantică sau intențională a lumii. Programele de AI generează răspunsuri deosebit de complexe, însă ele nu pot înțelege ce anume înseamnă acele răspunsuri. Învățarea prin recompense (*reinforcement learning*), deși presupune evaluarea comportamentelor prin

valori numerice pozitive sau negative, nu este echivalentă cu generarea de stări valențiate în cadrul acestor sisteme. Aceste numere nu poartă o încărcătură semantică pentru că sistemele nu dețin structuri care să permită înțelegere sau simțirea unor stări care se simt bine sau rău. Lucrarea de față plasează sub semnul întrebării termenul de *inteligență* din expresia *inteligență artificială*. Dacă inteligența presupune adaptabilitate, generalizare și înțelegerea lumii, atunci inteligența artificială doar mimează această trăsătură. Mai mult, dacă înțelegerea este legată de stări afective și homeostatice, dar acestea pot apărea doar cu condiția existenței unui corp cu nevoi și care este degradabil, atunci programele de AI nu vor deveni inteligente în lipsa acestor trăsături. Mai mult, dacă definim intenționalitatea drept orientarea deliberată către lume, atunci AI-ul are nevoie de un mediu prin care această interacțiune să poată avea loc. Implementarea de senzori ar rezolva, în teorie, doar jumătate din problemă, întrucât orientarea către lume reprezintă un proces bidirecțional, prin care o entitate modifică mediul din jurul său, iar existența în acest

mediu produce consecințe ce permit apariția sensului și a intenționalității. Fără această dimensiune a consecințelor trăite, programele AI nu pot depăși stadiul actual al procesării de date în mod nediferențiat și fără intenție.

Din perspectivă socială și etică, relevanța cercetării se regăsește în discuțiile despre interacțiunile oamenilor cu programele de inteligență artificială. Antropomorfizarea acestor sisteme, împreună cu supraestimarea capacităților poate avea efecte negative de ordin social și poate genera riscuri etice. Prin antropomorfizare, limitele tehnologice devin mai puțin evidente, însă, după cum vom sublinia, sistemele artificiale sunt susceptibile fenomenului de halucinare, care se referă la faptul că nu de puține ori răspunsurile generate conțin informații false sau chiar inventate, însă această tendință nu este una deliberată. Halucinațiile artificiale apar pentru că sistemele nu pot ști ce anume generează și, drept urmare, nu pot verifica veridicitatea outputurilor. De asemenea, din cauza setului de date pe care sunt antrenate, răspunsurile pot adesea să

conțină biasuri rasiale sau de gen, care pot afecta calitatea informațiilor oferite. În plus, folosirea acestor programe în domenii precum justiția sau medicina ridică probleme legate de responsabilitate. Dacă un utilizator urmează sfaturi medicale oferite de un LLM, cine răspunde atunci când sfaturile se dovedesc a fi periculoase? Atunci când le atribuim înțelegere, riscăm să ajungem să ne raportăm la sistemele artificiale ca și cum acestea ar fi oameni de știință, punând în seama lor responsabilitatea inovației. Fără o minte capabilă de a produce idei cu adevărat noi, contribuțiile științifice ale inteligenței artificiale sunt aproape inexistente. Faptul că aceste sisteme nu posedă o înțelegere a lumii și nici un sistem axiologic permite întrebuințarea capacităților computaționale în scopuri dăunătoare societății. Astfel, programele de AI pot fi folosite pentru a genera și răspândi *fake news*, *deepfake*-uri și pentru a manipula opinia publică în scopuri politice, economice sau sociale. Toate aceste aspecte evidențiază necesitatea reflecției asupra limitelor acestor sisteme și asupra modului în care ele ar trebui reglementate și integrate în societate.

Din perspectivă interdisciplinară, lucrarea de față concentrează trei domenii care au fost adesea cercetate separat, anume filosofia minții, neuroștiințele și tehnologia inteligenței artificiale. Astfel, filosofia minții pune accent pe importanța experienței subiective și pe natura intenționalității, neuroștiințele oferă dovezi privind rolul stărilor afective și homeostatice, iar domeniul dezvoltării inteligenței artificiale are în vedere construcția sistemelor predictive. Prin abordarea multidisciplinară, cercetarea analizează limitele actuale ale AI-ului și oferă un cadru conceptual care servește drept punct de plecare pentru construcția uneltelor care ar putea, în teorie, să prezinte proto-intenționalitate. Trebuie menționat că acest model nu este infailibil, dar subliniază nevoia unor structuri homeostatice și a unui corp vulnerabil în lipsa cărora sistemele artificiale nu vor putea fi intenționale și, prin urmare, nu vor putea avea înțelegerea datelor cu care operează.

Astfel, primele două capitole ale lucrării de față au rolul de familiariza cititorul cu ideile și

conceptele din sfera filosofiei minții, respectiv a inteligenței artificiale și de a pune în lumină principalele divergențe dintre perspectivele filosofice și câteva dintre limitele tehnice din sfera AI. În primul capitol vom aborda teoriile clasice cu privire la fenomenul conștiinței, cu accent pe dualism, funcționalism și emergentism și vom analiza perspectivele unor autori precum John Searle, Daniel Dennett, Ned Block, Hilary Putnam sau David Chalmers. Această analiză permite o perspectivă cuprinzătoare asupra complexității conștiinței și contribuie la implicațiile întrebării de cercetare. Al doilea capitol are rolul de a stabili contextul istoric de la scrierile lui Alan Turing până în prezent. Vom discuta despre cele două abordări principale în construcția sistemelor artificiale, anume cea simbolică și cea neuronală. Ne vom raporta la autori precum Stuart Russell, Peter Novig, Tom Michael Mitchell pentru analiza modului de funcționare al învățării automate și la Gary Marcus, care critică ideea că modelele bazate exclusiv pe date vor putea conduce la sisteme cu adevărat inteligente. În cel de-al treilea

capitol vom defini termenii de *conștiință* și *inteligență artificială* și vom stabili cadrul conceptual în care vom lucra pe parcursul cercetării.

Cea de-a doua parte a tezei se concentrează asupra mecanismelor cerebrale și afective care contribuie la fenomenul conștiinței. Vom analiza perspective precum teoria celor o mie de creieri a lui Jeff Hawkins, principiul energiei libere enunțat de Karl Friston și Mark Solms, procesarea predictivă. De asemenea, vom analiza critic cele două principale teorii cu privire la natura stărilor afective, anume teoria emoțiilor de bază (BET) a lui Solms și Panksepp și teoria emoțiilor construite (TCE), propusă de Lisa Feldman Barrett. Vom arăta de ce nici BET și nici TCE nu pot să ofere o perspectivă completă asupra emoțiilor. Prima ignoră substratul socio-cultural al afectelor iar cea de-a doua nu acordă suficientă importanță datului biologic. Apoi, ne vom orienta către perspectivele lui Joseph LeDoux și Antonio Damasio pentru a analiza modul în care stările afective sunt cele care transformă experiența în experiență trăită, resimțită ca *a mea*. De asemenea,

împreună cu stările homeostatice, afectele contribuie la modul în care atât lumea exterioară, cât și cea interioară capătă sens.

În a treia parte vom reveni asupra inteligenței artificiale și vom intra în detalii conceptuale privind rețelele de tip transformer care se află în spatele LLM-urilor precum ChatGPT. Tot aici vom analiza ideea conform căreia sistemele artificiale pot efectua cele trei tipuri de raționamente: deductive, inductive și abductive. Vom relua argumentele lui Gary Marcus și ale lui Erik Larson privind capacitatea acestor programe de a înțelege datele de procesat și ne vom orienta către abducție, așa cum a fost formulată de către Peirce, pentru a arăta că modelele actuale nu pot reproduce creativitatea umană de care este nevoie pentru a efectua raționamente abductive. De asemenea, vom integra sugestia lui Marcus cu privire la renunțarea la arhitectura bazată exclusiv în detrimentul unei abordări hibride, neuro-simbolice și vom extinde această sugestie pentru a propune câteva trăsături minim necesare pentru emergența proto-semnificației în cadrul sistemelor AI.

Ultimele capitole au în vedere perspective etice și sociale în ce privește utilizarea și viitorul inteligenței artificiale. Aici, vom discuta despre întrebuințarea acestor programe în scopuri dăunătoare societății, dar și despre implicațiile antropomorfismului. De asemenea, vom aminti scrierile lui Nick Bostrom, Ray Kurzweil, Eliezer Yudkowsky și Roman Yampolskiy. Acestea ne vor permite să analizăm critic scenariile speculative cu privire la viitorul AI-ului și pentru a sublinia cum astfel de discursuri distrag atenția de la riscurile pe care aceste sisteme le implică deja. În această parte, argumentul se conturează astfel: în lipsa intenționalității și a înțelegerii, inteligența artificială trebuie privită drept o unealtă puternică, dar nu trebuie să pierdem din vedere că este fundamental diferită de mintea umană și, ca atare, demersurile de a construi agenți autonomi AI nu ar trebui să urmărească înlocuirea acesteia. Evoluțiile din domeniul AI contribuie la trivializarea complexității minții umane. De fiecare dată când discursul public se concentrează pe cât de aproape suntem de dezvoltarea agenților AI,

se creează impresia că inteligența artificială generală este iminentă, ignorând limitele fundamentale ale acestor unelte. Această supraestimare distorsionează percepția publicului și riscă să creeze așteptări nerealiste. Dacă ne uităm la istoria inteligenței artificiale, promisiunile exagerate au condus la fenomenul numit *iarna inteligenței artificiale*, moment în care această tehnologie dispare din atenția publică din cauza dezamăgirii cu privire la capacitățile reale ale sistemelor AI.

Pe întreg parcursul lucrării vom apela la scrierile unor autori din domenii precum filosofie, inteligență artificială și neuroștiințe pentru a demonstra că conștiința și intenționalitatea nu pot să apară în lipsa stărilor afective și homeostatice cu miză și a unui imperativ de supraviețuire. Ca atare, orice sistem artificial care nu posedă aceste trăsături reale, nu simulate, nu poate avea pretenția inteligenței pe care o impune denumirea de inteligență artificială generală.

Motivația și relevanța temei alese

Alegerea acestei teme de cercetare a fost determinată de interesul pe care l-am manifestat față de acest domeniu pe parcursul ultimilor opt ani. Mai mult, contactul tot mai profund cu tehnologia a amplificat acest interes, iar proiectul inteligenței artificiale generale reprezintă tocmai puntea dintre problema conștiinței și tehnologie. Dincolo de motivațiile personale, cercetarea inteligenței artificiale prin prisma probleme conștiinței reprezintă un demers ce merită toată atenția. Problema conștiinței și cea a inteligenței artificiale au fost cercetate de-a lungul timpului, însă predominant în mod separat. Teoreticienii inteligenței artificiale provin, în general, din domenii tehnice precum inginerie sau robotică, însă aceștia rareori atacă acest concept din perspectivă filosofică. Pe de altă parte, textele cu privire la problema conștiinței au în vedere predominant aspecte ce țin de neuroștiințe sau filosofie. Există autori preocupați de conștiință care dedică secțiuni sau capitole din lucrările lor inteligenței artificiale, însă problema este abordată superficial sau prea puțin. Acest aspect a condus, în mod firesc, la motivația de a

cerceta granița dintre filosofie, neuroștiință și inteligență artificială.

În timp ce progresul tehnologic a condus la dezvoltarea unor sisteme artificiale din ce în ce mai performante, discuțiile despre inteligența înțeleasă ca adaptabilitate, generalizare și ca trăsătură care poate apărea doar în urma experienței trăite au rămas în plan secundar. Iluzia iminenței agenților AI, înzestrați cu conștiință și intenționalitate, împreună cu tendința umană de a antropomorfiza artefactele din jurul nostru conduc la supraestimarea capacităților sistemelor artificiale contemporane. În ciuda complexității lor, LLM-urile generează adesea răspunsuri eronate. Ca atare, relevanța temei alese rezidă în nevoia de a aduce în spațiul public și academic limitările sistemelor artificiale care, după cum va fi discutat, sunt considerabil mai numeroase decât ar părea la o primă interacțiune cu aceste programe. Mai mult, este necesar să ne întrebăm dacă cercetarea predominantă în domeniul inteligenței artificiale, cea bazată pe cantități tot mai mare de date și pe investiții masive în aceeași arhitectură reprezintă direcția corectă.

Progresele din ultimii ani par relativ modeste, comparativ cu salturile tehnologice realizate în urmă cu aproape un deceniu, fapt ce sugerează o stagnare. În acest context, relevanța cercetării vizează și impactul social și științific al acestor programe. Un risc este cel ca oamenii să ajungă să delege facultatea critică acestor sisteme, în ciuda faptului că ele nu înțeleg ce procesează și nu dispun de o dimensiune afectivă, intențională sau axiologică.

După cum subliniază Stuart Russell, încă este prea devreme pentru a discuta despre alinierea scopurilor inteligenței artificiale cu scopurile omenirii, ci, mai degrabă, discursurile ar trebui să abordeze teme precum alinierea intereselor oamenilor cu scopurile omenirii. Pentru că sistemele artificiale sunt departe de a poseda înțelegere, modele cauzale ale lumii, intenționalitate sau conștiință, una dintre problemele centrale este de a ne asigura că oamenii nu proiectează asupra AI-ului trăsături pe care nu le are.

Precauții și delimitări

Pe parcursul acestei cercetări, vom avea în vedere fenomenele conștiinței și intenționalității în cheie umană sau antropocentristă. Am ales acest cadru pentru că oamenii oferă cel mai clar exemplu al acestor două concepte și datorită literaturii de specialitate abundentă. Astfel, când vom discuta despre *intenționalitate* și *conștiință*, ne vom referi la manifestările lor umane. Trebuie menționat faptul că nu excludem posibilitatea ca intenționalitatea, inteligența sau conștiința să apară în forme de organizare care nu reproduc mecanismele umane. Așadar, este posibil ca, în viitor, să apară sisteme artificiale intenționale și conștiente, ale căror stări derivă din funcționalități sau mecanisme complet diferite de cele care permit apariția acestor trăsături în cazul oamenilor. Mai mult, nu excludem emergența acestor trăsături în forme non-umane. Totuși, această teză se raportează la stadiul cercetărilor contemporane și va discuta despre aceste trăsături artificiale pornind de la intenționalitatea și conștiința umană.

Un aspect relevant este și cel al intenționalității umane. Literatura de specialitate (Searle, Dennett)

distinge între două tipuri de intenționalitate: cea originară (sau intrinsecă) și cea derivată. Pentru Searle, oamenii posedă intenționalitate originară, iar sistemele artificiale au doar intenționalitate derivată. În *Evolution, Error and Intentionality*, Dennett argumentează că intenționalitatea umană este, de fapt, derivată, pentru că ceea ce numim intenționalitate originară provine din imperativele de supraviețuire sau din datul biologic. Astfel, intenționalitatea este un aspect problematic, dar vom pleca de la ideea că, în forma sa umană, intenționalitatea este originară și experiențială. Ca atare, vom cerceta dacă sistemele de inteligență artificială ar putea avea intenționalitate intrinsecă, similară cu cea umană.

Metodologie

Lucrarea de față se bazează pe o analiză filosofică și conceptuală a raportului dintre conștiință și inteligență artificială. Natura întrebării de cercetare (*pot sistemele artificiale să fie conștiente și intenționale?*) necesită clarificare conceptuală și o analiză critică a teoriilor existente. Primul pas constă

în clarificarea conceptelor de bază, anume *conștiința*, *intenționalitate*, *afecte*, *semnificație*, *inteligență artificială*. Vom examina aceste concepte atât din perspectivă filosofică (Searle, Dennett, Chalmers), cât și din perspectiva neuroștiințelor (Friston, Solms, Barrett, LeDoux), dar și a cercetătorilor din domeniul inteligenței artificiale (Marcus, Yampolskiy, Mitchell). Prin acest demers, lucrarea va urmări evitarea ambiguității și crearea unui cadru conceptual clar. De asemenea, vom folosi analiza comparativă pentru a distinge între abordările din domeniul AI (de la arhitecturi simbolice la cele neuronale). Mai mult, vor fi analizate și teoriile filosofice și neuroștiințifice cu privire la conștiință pentru a sublinia ce anume lipsește din construcția sistemelor artificiale pentru ca acestea să poată avea experiențe subiective și intenționalitate. De asemenea, vom analiza critic și încercările de formalizare ale raționamentelor logice cu scopul de a evidenția lipsa dimensiunii creative a acestor programe.

Mai departe, vom integra rezultatele cercetărilor din trei domenii distincte pentru a avea o

perspectivă amplă, multidisciplinară atât asupra fenomenului conștiinței, cât și asupra posibilității ca sistemele artificiale să dezvolte în viitor trăsături precum stări afective, nevoi corporale cu miză sau intenționalitate. În ultimul rând, lucrarea va propune un model conceptual minimal, cel al proto-intenționalității, care poate servi drept punct de plecare pentru discuțiile și cercetările ce vizează posibilitatea dezvoltării *bottom-up* a viitoarelor sisteme artificiale. Acest model nu are pretenția de a fi complet și reprezintă mai degrabă o perspectivă conceptuală care poate fi utilă pentru explorarea intenționalității artificiale.

Pe parcursul cercetării doctorale, rezultatele parțiale au fost prezentate în cadrul a două conferințe academice, iar o unele dintre concluziile cercetării au fost publicate sub forma unui articol științific. Secțiunile marcate cu asterisc (*) reprezintă aspecte prezentate anterior în cadrul conferințelor, în lucrarea de disertație sau în articolul științific, însă în prezenta lucrare ele sunt dezvoltate într-o versiune mai amplă.

În 2016, Peter Brown a publicat cartea pentru copii *The Wild Robot* (*Robotul sălbatic*). Povestea are în prim plan un robot pe nume Rozzum Unit 7134 sau, pe scurt, Roz, care este activat din greșeală în urma unui naufragiu pe o insulă nelocuită. Roz se trezește singură, în natură. La început, se chinuie să supraviețuiască, pentru că nu a fost concepută pentru a se descurca în sălbăticie, iar animalele o privesc ca pe o amenințare. Dar Roz se adaptează și începe să înțeleagă limbajul animalelor. Apoi, robotul adoptă un pui de gâscă numit Birghtbill, iar astfel se integrează și mai bine în lumea animalelor. Datorită bunătății sale și a dorinței de a învăța, Roz se împrietenește cu animalele și devine membru al comunității. Totuși, Roz nu poate ignora natura sa, astfel încât creatorii ei se întorc pentru a o recupera și a o plasa în rolul pentru care a fost construită, anume cel de asistent pentru oameni. Povestea se încheie cu robotul părăsind insula, dar cu dorința de a se întoarce pe insulă. Roz a fost creată pentru a urma instrucțiunile oamenilor și pentru a le spori productivitatea, întocmai scopul pentru care unii cercetători încearcă să

construiască agenți AI. Plasată într-un context complet diferit de cel pentru care a fost creată, Roz trebuie să învețe și să se adapteze pentru a-și asigura supraviețuirea. Transformarea ei dintr-o unealtă într-un membru al comunității animalelor sălbatice ilustrează ideea că semnificațiile pot apărea doar prin interacțiune și conexiune afectivă cu lumea. Roz capătă intenționalitate datorită experienței trăite și a interacțiunilor cu lumea, în scopul supraviețuirii.

Povestea subliniază argumentele expuse în lucrarea de față, anume faptul că intenționalitatea și conștiința nu pot fi reduse la computații bazate pe analiza probabilistică a unor informații procesate nediferențiat. Aceste două elemente necesită mize, nevoi și posibilitatea eșecului sub forma degradării și/sau a morții, la fel cum Roz nu ar fi putut deveni cine este fără a se confrunta cu pericole și fără a acționa în conformitate cu stările afective. Spre deosebire de Roz, inteligența artificială contemporană nu are un corp, o lume semantică sau stări afective și nu poate suferi. Povestea subliniază că mașina trebuie să treacă dincolo de procesare informațională pentru a

fi mai mult decât o unealtă. Pentru a deveni o entitate intenţională cu experienţe trăite, inteligenţa artificială are nevoie de un corp, de mize şi imperative de supravieţuire, manifestate prin stări afective şi homeostatice.

Robotul sălbatic evidenţiază motivele pentru care inteligenţa artificială în forma sa contemporană este complet diferită de ceea ce înţelegem prin inteligenţă umană. Deşi vorbim despre inteligenţă, lucrarea de faţă a avut drept punct de plecare conştiinţa şi intenţionalitatea, dar, de fapt, niciuna dintre aceste două trăsături nu pot apărea în cadrul sistemelor artificiale aşa cum sunt ele construite astăzi. Fără conştiinţă nu există cunoaştere de sine, iar fără cunoaştere de sine nu putem vorbi despre gândire. La fel, fără intenţionalitate, adică fără orientarea către lume şi fără posibilitatea de a interacţiona cu aceasta, nu există inteligenţă. Întrebarea centrală a lucrării de faţă este dacă putem construi sisteme artificiale cu conştiinţă şi intenţionalitate. Pentru a răspunde, am pornit de la cel mai intim fenomen pentru fiecare dintre noi – conştiinţa. Am analizat principalele teorii

și am arătat de ce multe dintre ele sunt insuficiente pentru a explica acest fenomen. Spre exemplu, funcționalismul poate fi privit drept o altă formă a behaviorismului, unde contează doar funcțiile și outputurile unui sistem, iar stările interne sunt ignorate. Din această cauză, funcționalismul nu poate explica experiențele trăite, care sunt inseparabile de fenomenul conștiinței. Am argumentat că abordările cogniției întrupate sunt mai promițătoare. Apoi, am conturat un scurt istoric al inteligenței artificiale, pentru a contextualiza discuția și pentru a familiariza cititorul cu principalele teme din acest domeniu. Tot aici am subliniat limitările curente și am argumentat că sistemele artificiale funcționează optim în sisteme închise, însă inteligența presupune capacitatea de adaptare, de generalizare și de abstractizare.

Ulterior, am argumentat cadrul teoretic în care am lucrat, anume emergentismul slab ca potențială direcție de cercetare pentru a explica natura conștiinței, pe care am analizat-o în triada conștiință-corp-afecte. Fără corp nu există intenționalitate, iar fără afecte nu există experiență

trăită. Experiența trăită reprezintă esența conștiinței, ca flux mental de imagini pe care le resimt *ale mele*. Am argumentat că trebuie să privim conștiința ca fenomen întrupat, dependent de deviații homeostatice manifestate prin stări afective valențiate. Apoi am prezentat modelul mecanismului de procesare predictivă, orientat spre reducerea energiei libere înțeleasă ca entropie, precum și conceptul de zonă Markov ca posibilă funcție a conștiinței. Concret, construim modele interne care se modifică în funcție de interacțiunile noastre în lume, iar scopul procesării predictive este de a anticipa momentele în care realitatea nu coincide cu predicțiile noastre. Deși aceste teorii nu sunt unanim acceptate, ele oferă un cadru conceptual util pentru înțelegerea relației dintre predicție, adaptare și experiență. Am analizat apoi cele două mari teorii cu privire la emoții, anume teoria emoțiilor de bază (BET) și teoria emoțiilor construite (TCE). Prima susține existența a șapte emoții de bază universale și înnăscute, însă nu s-au putut identifica localizări precise care să confirme această teorie. Cea de-a doua teorie (TCE) susține că emoțiile sunt

fenomene socio-culturale, însă această abordare ignoră importanța dimensiunii biologice. Am subliniat că niciuna dintre aceste teorii nu poate explica în totalitate natura emoțiilor. În schimb, am insistat asupra rolului fundamental al afectelor, întrucât ele fac ca experiența să fie recunoscută imediat drept experiență proprie și sunt cele care creează sentimentul de posesie asupra fluxului mental. Ca atare, pentru că fenomenul conștiinței este unul personal, recunoscut întotdeauna ca *al meu*, afectele trebuie să reprezinte elemente constitutive ale acestui fenomen.

Am trecut apoi la analiza sistemelor artificiale și am descris conceptual modul în care acestea funcționează, cu accent pe arhitectura transformer, care se regăsește în majoritatea programelor lingvistice disponibile astăzi. Am investigat în ce măsură pot opera pe baza inferențelor logice. Din punct de vedere deductiv, sistemele artificiale pot funcționa, doar că deducția nu conduce la noi informații. În ce privește inducția, AI-ul generalizează cu succes de multe ori în seturi limitate de informații,

însă procesul este constrâns de datele limitate pe baza cărora a fost antrenat. Când vine vorba de abducție, am demonstrat că, în ciuda încercărilor, inferențele bayesiene nu pot reprezenta formalizarea completă a inferențelor abductive, pentru că acestea presupun generarea de ipoteze noi printr-un proces deseori creativ, iar inferențele bayesiene presupun doar selectarea celei mai probabile explicații dintr-o serie de ipoteze predefinite. Ne-am orientat apoi către limitările tehnologice actuale, printre care amintim ipoteza frecvenței, constrângerea empirică și, poate cea mai relevantă dintre ele, saturația modelelor. Această limitare se referă la faptul că modelele nu pot fi antrenate la nesfârșit și că, la un moment dat, informațiile noi nu mai pot îmbunătăți capacitățile AI-ului, nici nu pot altera informațiile preexistente. Am analizat și problema creativității în cadrul sistemelor artificiale și am argumentat că este o iluzie, întrucât simpla reordonare simbolică neintenționalăă nu poate constitui un act creativ. Creativitatea presupune intenționalitate și exprimare afectivă. În artă, spre exemplu, actul creativ poartă rol de reglare afectivă, la

fel cum medicina poate fi considerată o știință dezvoltată cu scopul de a evita moartea, adică imperativul biologic care conferă baza necesară pentru apariția semnificațiilor.

În continuare, am arătat că inteligența artificială nu poate fi intențională, întrucât nu are nevoi care să necesite remediere, iar intenționalitatea presupune orientarea către lume într-un proces bidirecțional, unde acțiunile noastre în lume au consecințe atât interioare, cât și exterioare. Chiar dacă am simula nevoi cu scopul de a permite apariția unei proto-semnificație, de exemplu nevoia de a economisi energie sau nevoia de a acumula date, acestea nu contează pentru sistemele artificiale, pentru că nu produc modificări cu miză. Semnificația apare doar atunci când există consecințe reale, iar nevoile simulate în calitate de risc implementat extern nu au nicio miză. Foamea și amenințarea morții poartă o încărcătură semantică tocmai pentru că sunt resimțite. În următoarea secțiune am analizat câteva teorii care susțin că inteligența artificială bazată pe date ar putea manifesta un fel de conștiință și am arătat de ce aceste

abordări sunt eronate, apoi am discutat despre *reinforcement learning* în calitate de potențială abordare pentru a genera stări valențiate în cadrul sistemelor artificiale, dar am argumentat că recompensele și pedepsele acordate acestor sisteme nu sunt resimțite, scopul valorilor numerice fiind de a ghida comportamentele ulterioare. Pe de altă parte, în cazul oamenilor, comportamentul este ghidat de mecanisme predictive, de homeostază și de stări afective valențiate. Pe parcursul secțiunii III.3 am formulat condițiile minim necesare pentru emergența intenționalității în sistemele artificiale: 1) existența unui corp, un imperativ de supraviețuire care să guverneze deciziile, 2) a senzorilor capabili să măsoare variabile reale și 3) a homeostazei cu rol de transmitere a acestor deviații într-o formă afectivă cu valență. Mai mult, am susținut necesitatea unei arhitecturi neuro-simbolice, unde partea simbolică ar permite implementarea regulilor și, eventual, a simțului comun, iar partea neuronală ar facilita procesul de identificare a tiparelor din informații noi. Nu știm dacă această abordare ar conduce la AGI, dar

ar putea, în teorie, reduce halucinațiilor modelelor bazate exclusiv pe date. În cadrul acestui potențial model, deviațiile de la normă ar trebui identificate drept pericole. Pentru a fi considerat agent AI, sistemul ar trebui să comporte o dimensiune afectivă sub forma deviațiilor resimțite drept amenințări. Această idee este susținută de cercetările recente care arată că întruparea este o condiție necesară în cazul construcției agenților artificiali, după cum am arătat în secțiunea III.3.4.

Modelul pe care l-am propus este limitat, însă arată că abordările actuale nu ne vor aduce mai aproape de proiectul AGI. În lipsa unor mecanisme care să implice modele interne ale lumii, imperative de supraviețuire şi module homeostatice care monitorizează deviații, nu putem discuta despre dezvoltarea unui sistem artificial care să posede inteligență de tip uman, conștiință sau intenționalitate. Ulterior ne-am îndreptat atenția către riscurile iminente ale inteligenței artificiale. Un fenomen esențial care contribuie la aceste riscuri este antropomorfismul, adică tendința de a ne raporta la

sisteme artificiale ca și cum ar fi intenționale, ar comporta o dimensiune afectivă sau ar poseda cunoștințe explicite. Această tendință se datorează faptului că avem tendința de a ne extinde intenționalitatea și de a o proiecta asupra artefactelor din jurul nostru, conform filosofului Searle. Prin folosirea verbelor de acțiune specific umane de tipul *AI-ul știe* sau *înțelege*, ajungem să le investim cu expertiză pe care nu o au. Această proiecție conduce, în timp, la impresia că aceste sisteme ar înțelege outputurile pe care le generează. Am analizat și utilizarea AI-ului în scopuri malițioase sau chiar ilegale: *deepfake*-uri, campanii de dezinformare și *cybercrime*. De asemenea, am subliniat necesitatea reglementărilor clare.

În capitolul IV.2 am discutat despre scenariile speculative legate de superinteligență și am argumentat că ele supraestimează capacitățile actuale și mută discursul public într-o direcție alarmistă, care nu are corespondent în realitatea tehnologică. De asemenea, am arătat că o parte dintre inginerii care dezvoltă aceste sisteme adoptă o perspectivă

behavioristă, confundând outputurile cu competență. În ultimul capitol, am subliniat rolul culturii asupra inteligenței umane. Cultura este un fenomen constitutiv al inteligenței și presupune participare colectivă și semnificații împărtășite. Deoarece nu are acces la dimensiunea semantică, subiectivă și intențională necesară, AI-ul nu poate participa la cultură. În acest sens, inteligența este un fenomen inerent social, dependent de interacțiunea cu celălalt. Am avut în vedere și impactul social al utilizării acestor sisteme, atât în termeni culturali, dar și în ce privește percepția publică asupra științei. În final, am adus în discuție ipoteza unei *noi ierni a inteligenței artificiale*, când promisiunile legate de AGI nu se materializează, iar termenul dispare temporar din atenția publică. În mod paradoxal, această diminuare a entuziasmului ar putea fi benefică, pentru că ar putea permite progreselor să aibă loc într-un climat mai puțin dominat de agitație și de promisiuni făcute sub presiunea inovației cât mai rapide.

Contribuții personale

Lucrarea de față propune mai multe perspective originale în ce privește raportul dintre conștiință și inteligență artificială. În primul rând, am argumentat că afectele sunt constitutive pentru fenomenul conștiinței, fiind elemente care conferă experienței caracterul subiectiv. Astfel, am plasat stările afective în centrul teoriei asupra conștiinței. În al doilea rând, am susținut că intenționalitatea necesită existența nevoilor reale care apar datorită imperativului de supraviețuire și a homeostazei. Aceste nevoi trebuie să aibă o miză care să fie resimțită sub forma afectelor. Nevoile simulate nu pot conduce la semnificații pentru că nu produc consecințe resimțite. În al treilea rând, am argumentat că inferențele bayesiene, deși pot fi translate în limbaj computațional, nu reprezintă o formalizare completă inferențelor abductive. În aceeași manieră, am subliniat dimensiunea creativă necesară pentru abducție și am argumentat că sistemele artificiale nu sunt capabile de acte creative, deoarece creativitatea necesită intenție și expresia stărilor afective. De

asemenea, am introdus conceptul de *behaviorism 2.0*, care descrie tendința contemporană de a confunda outputurile sistemelor artificiale cu competența sau inteligența, în același fel în care behavioriști considerau comportamentele drept rațiuni suficiente pentru a afirma stările mentale subiective ale oamenilor. Mai mult, am folosit procesarea predictivă și principiul energiei libere pentru a descrie mecanismele predictive cerebrale, care au rolul de a prezice și a corecta erorile de procesare sau de a minimiza nivelul de energie liberă din organism. Împreună cu homeostaza și stările afective, am argumentat că aceste elemente sunt indispensabile pentru emergența conștiinței.

Teza propune un model conceptual minim pentru apariția proto-semnificației în cadrul sistemelor artificiale, bazat pe patru condiții: 1) un corp fizic; 2) imperative de supraviețuire; 3) senzori care să măsoare buna funcționare a sistemului și 4) o arhitectură hibridă (neuro-simbolică). Acest model nu este unul complet, ci reprezintă mai degrabă o direcție viitoare de cercetare, având în vedere neajunsurile

sporirii capacităților computaționale în cadrul sistemelor artificiale bazate pe date. Altfel spus, indiferent de cât de sofisticate par a fi răspunsurile generate de programele de AI contemporane, ele nu au acces la dimensiunea semantică a informațiilor procesate, nu sunt intenționale și nu au nevoi interne sau stări afective. Aceste aspecte conduc la concluzia principală a cercetării de față. În lipsa unui corp fizic, a nevoilor, a imperativelor de supraviețuire și a stărilor afective, inteligența artificială rămân programe predictive. Pentru a înfăptui proiectul AGI sau pentru a construi agenți AI, este improbabil ca abordarea bazată exclusiv pe procesarea nediferențiată a datelor să conducă la rezultatele așteptate.

Direcții viitoare de cercetare

Contribuțiile expuse anterior conturează câteva direcții viitoare de cercetare. În primul rând, este necesară aprofundarea rolului pe care stările afective îl au în cadrul fenomenului conștiinței și investigarea modului în care acestea ar putea fi implementate sau generate în sistemele artificiale. Astfel, este important

să răspundem la întrebarea *pot fi reduse stările afective la mecanisme neuronale sau este nevoie de întrupare cu nevoi și mize?* În al doilea rând, condițiile minim necesare pentru construcția sistemelor proto-intenționalitate pot fi detaliate și testate prin proiecte interdisciplinare, la intersecția dintre robotică, neuroștiințe și filosofia minții. Aceste experimente ar putea investiga în ce măsură putem construi mașini intenționale. O altă potențială direcție de cercetare este redefinirea ideii de inteligența artificială. Conceptul de *behaviorism 2.0* poate reprezenta un instrument critic pentru a distinge între competență simulată și reală, care poate contribui la reevaluarea performanțelor sistemelor AI. Nu în ultimul rând, o posibilă direcție este cea a integrării mecanismelor predictive și homeostatice în cadrul sistemelor artificiale, pentru a testa dacă aceste mecanisme ar putea, în teorie, să apropie aceste sisteme de o formă de înțelegere semantică. Arhitectura hibridă, după cum am argumentat, ar putea, în teorie, să permită crearea unui model intern descriptiv al lumii, iar acest model ar putea fi compatibil cu procesarea predictivă. De

asemenea, modul în care utilizăm inteligența artificială în forma sa contemporană reprezintă un aspect care trebuie analizat, pentru a limita pe cât posibil impactul negativ pe care aceste unelte îl au asupra societății, culturii și psihicului uman.

Lucrarea de față subliniază că inteligența, conștiința și intenționalitatea nu pot fi separate de condițiile care permit apariția lor, anume stări afective, nevoi corporale și mize. Fără aceste lucruri, sistemele artificiale vor rămâne programe predictive care nu au acces la dimensiunea semantică și afectivă a lumii, care este condiția necesară pentru inteligență și, prin extrapolare, pentru emergența conștiinței. Această perspectivă conturează ideea că trăsăturile pe care le asociem cu fenomenul conștiinței se află în strânsă legătură cu experiențele trăite și simțite ca fiind *ale mele*, iar aceste experiențe comportă adesea o dimensiune socio-culturală. Astfel, sistemele artificiale nu trebuie privite ca entități care pot procesa informațiile atât sintactic, cât și semantic. În lipsa altor modificări arhitecturale, ele rămân doar unelte folositoare în domenii specifice, însă orice

întrebuințare trebuie să aibă în vedere limitările conceptuale și tehnologice de care dau dovadă aceste programe. Prin propunerea condițiilor minim necesare pentru apariția proto-intenționalității și prin sublinierea importanței stărilor afective și homeostatice, cercetarea de față deschide direcții de cercetare interdisciplinare. Înainte de a întreprinde un asemenea demers, ar trebui să ne întrebăm, dincolo de scenariile alarmiste cu privire la viitorul inteligenței artificiale, dacă ar trebui să încercăm construcția unor agenți AI, având în vedere că riscurile par să fie în număr mai mare decât beneficiile.

În același timp, indiferent de forma sau arhitectura acestora, viitoarele sisteme artificiale nu trebuie să înlocuiască ingeniozitatea minții umane, unde actul rațional și creativ produce idei noi de fiecare dată. De asemenea, atragem atenția asupra efectelor pe care AI-ului în forma prezentă le are în ce privește gândirea critică. Este o greșeală ca această facultate cognitivă să fie delegată unor programe care pot adesea să halucineze și care nu își pot da seama când generează răspunsuri eronate. Sub aceeași formă,

cercetarea științifică trebuie să prioritizeze contribuțiile umane în detrimentul rezultatelor obținute prin intermediul sistemelor artificiale. Prin aceasta, ne referim la faptul că cercetătorii nu trebuie să devină curatori de date pentru AI, iar rezultatele generate în laborator de către aceste programe trebuie mereu interpretate de o minte umană.

Evoluția lui Roz din *The Wild Robot* ilustrează argumentele principale ale lucrării de față. Conștiința și intenționalitatea pot apărea doar atunci când cogniția este însoțită de afecte, și un corp cu nevoi resimțite. Inteligența artificială contemporană nu beneficiază de aceste elemente constitutive și nu poate fi intențională în sens filosofic. De asemenea, AI-ul nu poate fi conștient și, deci, nu poate participa la dimensiunea semantică a lumii. Ca atare, orice manifestare care poate părea inteligentă sau intențională a programelor artificiale este doar rezultatul imensei puteri computaționale și a diversității artefactelor culturale dezvoltate de omenire de-a lungul anilor. Povestea lui Roz ne arată că supraviețuirea, stările afective și legătura cu lumea

exterioară sunt necesare pentru afirmarea intenționalității sau a conștiinței în cadrul oricărui sistem care poate accesa dimensiunea semantică. Fără aceste trăsături, sistemele artificiale vor rămâne unelte și nu agenți.

Bibliografie

Adnan, Al; Rahman, Saifur; Maliha, Jeba; Shahrear, Shahjada; Tinny, Sejuti Sarker; Mohammad, Khalid; Haider, Syed Ali; „Analysis of Identification of Cybercrimes Using Cyber Security Analytics Powered by Artificial Intelligence", în *Chinese Science Bulletin*, vol. 69, nr. 7, 2024, pp. 2595–2604.

Arkoudas, Konstantine, „ChatGPT is no Stochastic Parrot. But it also Claims that 1 is Greater than 1", în *Philosophy & Technology*, vol. 36, nr. 3, 2023, pp. 54-1–54-29.

Arulkumaran, Kai; Deisenroth, Marc Peter; Brundage, Miles; Bharath, Anil Anthony; „Deep Reinforcement Learning: A Brief Survey", în *IEEE Signal Processing Magazine*, vol. 34, nr. 6, 2017, pp. 26–38.

Asada, Minoru, „Artificial Pain May Induce Empathy, Morality, and Ethics in the Conscious Mind of Robots", în *Philosophies*, vol. 4, nr. 38, 2019.

Asada, Minoru; Nagai, Yukie; Ishihara, Hisashi; „Why Not Artificial Sympathy?", în *Proceedings of the International Conference on Social Robotic*s, 2012, pp. 278–287.

Ascher, Marcia, „A River-Crossing Problem in Cross-Cultural Perspective”, în *Mathematics Magazine*, vol. 63, nr. 1, 1990, pp. 26–29.

Baars, Bernard J., „Global workspace theory of consciousness: toward a cognitive neuroscience of human experience”, în *Progress in Brain Research*, vol. 150, 2005, pp. 45–53.

Bahdanau, Dzmitry; Cho, Kyunghyun; Bengio, Yoshua; „Neural Machine Translation by Jointly Learning to Align and Translate”, la *International Conference on Learning Representations*, 2015.

Bansal, Gaurang; Chamola, Vinay; Hussain, Amir; Guizani, Mohsen; Niyato, Dusit; „Transforming Conversations with AI—A Comprehensive Study of ChatGPT”, în *Cognitive Computation*, vol. 16, 2024, pp. 2487–2510.

Barrett, Lisa Feldman, „Are emotions natural kinds?”, în *Perspectives on Psychological Science*, vol. 1, nr. 1, 2006, pp. 28–58.

Barrett, Lisa Feldman, „Emotions are real”, în *Emotion*, vol. 12, nr. 3, 2012, pp. 413–429.

Barrett, Lisa Feldman, „The theory of constructed emotion: an active inference account of interoception and

categorization”, în *Social Cognitive and Affective Neuroscience*, vol. 12, nr. 1, 2017, pp. 1–23.

Barrett, Lisa Feldman, *How Emotions Are Made: The Secret Life of the Brain*, Pan Macmillan, Londra, 2018.

Barrett, Lisa Feldman; Lindquist, Kristen A.; Bliss-Moreau, Eliza; Duncan, Seth; Gendron, Maria; Mize, Jennifer; Brennan, Lauren, „Of Mice and Men: Natural Kinds of Emotions in the Mammalian Brain? A Response to Panksepp and Izard”, în *Perspectives on Psychological Science*, vol. 2, nr. 3, 2007, pp. 297–311.

Barrett, Lisa Feldman; Simmons, W. Kyle, „Interoceptive predictions in the brain”, în *Nature Reviews Neuroscience*, vol. 16, 2015, pp. 419–429.

Berent, Iris; Vaknin, Vered; Marcus, Gary; „Roots, stems, and the universality of lexical representations: Evidence from Hebrew”, în *Cognition*, vol. 104, nr. 2, 2007, pp. 254–286.

Berrar, Daniel, „Bayes’ Theorem and Naive Bayes Classifier”, în *Encyclopedia of Bioinformatics and Computational Biology (Second Edition)*, vol. 1, 2025, pp. 483–494, preprint.

Bian, Ning; Han, Xianpei; Sun, Le; Lin, Hongyu; Lu, Yaojie; He, Ben; Jiang, Shanshan; Dong, Bin; „ChatGPT is a Knowledgeable but Inexperienced Solver:

An Investigation of Commonsense Problem in Large Language Models", 2023, https://arxiv.org/abs/2303.16421, accesat la 20.04.2025.

Block, Ned, „Troubles with functionalism", în *Minnesota Studies in the Philosophy of Science*, 1978, vol. 9, pp. 261–325.

Block, Ned, „Functionalism", în *Studies in Logic and the Foundations of Mathematics*, vol. 104, 1982, pp. 519–539.

Block, Ned, „Consciousness, accessibility, and the mesh between psychology and neuroscience", în *Behavioral and Brain Sciences*, vol. 30, 2007, pp. 481–548.

Bontridde, Noémi; Poullet, Yves; „The role of artificial intelligence in disinformation", în *Data & Policy*, vol. 3, 2021, pp. 32-1–32-21.

Borji, Ali, „A Categorical Archive of ChatGPT Failures", 2023, https://arxiv.org/abs/2302.03494, accesat la 20.04.2025..

Bostrom, Nick, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford, 2014.

Bostrom, Nick, „When Machines Outsmart Humans", în *Futures*, vol. 35, nr. 7, pp. 759–764.

Botvinick, Matthew; Wang, Jane X.; Dabney, Will; Miller, Kevin J.; Kurth-Nelson, Zeb; „Deep Reinforcement

Learning and Its Neuroscientific Implications", în *Neuron*, vol. 107, nr. 4, 2020, pp. 603–616.

Boyles, Robert James, „Artificial Qualia, Intentional Systems and Machine Consciousness", în *Proceedings of the Research@DLSU Congress 2012: Science and Technology Conference*, Filipine, 2012, pp. 110a–110c.

Brown, Matt; Klepper, David; „Fake images made to show Trump with Black supporters highlight concerns around AI and elections", în *Associated Press*, martie 2024, https://apnews.com/article/deepfake-trump-ai-biden-tiktok-72194f59823037391b3888a1720ba7c2, accesat la 11.06.2025.

Brown, Tom B.; Mann, Benjamin; Ryder, Nick; Subbiah, Melanie; Kaplan, Jared; Dhariwal, Prafulla; Neelakantan, Arvind; Shyam, Pranav; Sastry, Girish; Askell, Amanda; Agarwal, Sandhini; Herbert-Voss, Ariel; Krueger, Gretchen; Henighan, Tom; Child, Rewon; Ramesh, Aditya; Ziegler, Daniel M.; Wu, Jeffrey; Winter, Clemens; Hesse, Christopher; Chen, Mark; Sigler, Eric; Litwin, Mateusz; Gray, Scott; Chess, Benjamin; Clark, Jack; Berner, Christopher; McCandlish, Sam; Radford, Alec; Sutskever, Ilya; Amodei, Dario; „Language Models

are Few-Shot Learners", în *34th Conference on Neural Information Processing Systems*, Canada, 2020.

Bruiger, Dan, „Reflections on the AI alignment problem", în *AI & Society*, vol. 40, nr. 6, 2025, pp. 4383–4392.

Buckner, Cameron, „Deep learning: A philosophical introduction", în *Philosophy Compass*, nr. 14, 2019.

Chadha, Anupama; Kumar, Vaibhav; Kashyap, Sonu; Gupta, Mayank; „Deepfake: An Overview", în P.K. Singh, S.T. Wierzchoń, S. Tanwar, M. Ganzha, J.J.P.C. Rodrigues (eds.), *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security. Lecture Notes in Networks and Systems*, vol. 203, Springer, Dordrecht, 2021, pp. 557–566.

Chalmers, David, *The Conscious Mind*, Oxford University Press, New York, 1996.

Chalmers, David, „Facing up to the problem of consciousness", în *Journal of Consciousness Studies*, vol. 2, nr. 3, 1995.

Chen, Clarence, „The Mind-Body Problem: A Critique of Type Identity Theory", în *The Schola*, vol. 7, nr. 4, 2023, pp. 21–40.

Chen, Heather; Magramo, Kathleen; „Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’”, în *CNN*, februarie 2024, https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk, accesat la 11.06.2025.

Chen, J. X., „The Evolution of Computing: AlphaGo”, în *Computing in Science & Engineering*, vol. 18, nr. 4, 2016, pp. 4–7.

Chollet, François, „On the Measure of Intelligence”, 2019, https://arxiv.org/pdf/1911.01547, accesat la 17.02.2025.

Chomsky, Noam; Skinner, B. F.; „Verbal behavior”, în *Language*, vol. 35, nr. 1, 1959, pp. 26–58.

Chua, John Moses A., „Formulating Consciousness: A Comparative Analysis of Searle’s and Dennett’s Theory of Consciousness”, în *TALISIK*, vol. 4, 2017, pp. 43–62.

Churchland, Paul M., „Eliminative Materialism and the Propositional Attitudes”, în *The Journal of Philosophy*, vol. 78, nr. 2, 1981, pp. 67–90.

Clarke, Roger, „Asimov's Laws of Robotics: Implications for Information Technology”, în Michael Anderson, Susan Leigh Anderson (eds.), *Machine Ethics,* Cambridge University Press, Cambridge, 2011.

Cohen, Fred, „Computer viruses”, în *Computers & Security*, vol. 6, nr. 1, 1987, pp. 22–35.

Cohen, Michael A.; Cavanagh, Patrick; Chun, Marvin M.; Nakayama, Ken; „The attentional requirements of consciousness”, în *Trends in Cognitive Sciences*, vol. 16, nr. 8, 2012, pp. 411–417.

Cole, David, „The Chinese Room Argument”, în Edward N. Zalta, Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, 2024, https://plato.stanford.edu/archives/win2024/entries/chinese-room/, accesat la 25.04.2025.

Creed, Torrey; Salama, Leah; Slevin, Roisin; Tanana, Michael; Imel, Zac; Narayanan, Shrikanth; Atkins, David; „Enhancing the quality of cognitive behavioral therapy in community mental health through artificial intelligence generated fidelity feedback (Project AFFECT): a study protocol”, în *BMC Health Services Research*, vol. 22, 2022.

Crevier, Daniel, *AI: The Tumultuous Search for Artificial Intelligence*, BasicBooks, New York, 1993.

Cunningham, Padraig; Cord, Mathieu; Delany, Sarah Jane; „Supervised Learning”, în M. Cord, P. Cunningham (eds.), *Machine Learning Techniques for Multimedia*, Springer, Berlin, 2008, pp. 21–49.

Dalgleish, Tim; Dunn, Barnaby D.; Mobbs, Dean; „Affective neuroscience: Past, present, and future”, în *Emotion Review*, vol. 1, nr. 4, 2009, pp. 355–368.

Damasio, Antonio, *Strania ordine a lucrurilor: viața, sentimentele și nașterea culturilor*, trad. Carmen Strungaru, Humanitas, București, 2022.

Damasio, Antonio, *Simțire și cunoaștere: cum să generezi minți conștiente*, trad. Alexandru Babeș, Humanitas, București, 2024.

Das, Sumanjit; Nayak, Tapaswini; „Impact of Cyber Crime: Issues and Challenges”, în *International Journal of Engineering Sciences & Emerging Technologies*, vol. 6, nr. 2, 2013, pp. 142–153.

Das, Sumit; Dey, Aritra; Pal, Akash; Roy, Nabamita; „Applications of Artificial Intelligence in Machine Learning: Review and Prospect”, în *International Journal of Computer Applications*, vol. 115, 2015, pp. 31–41.

Dehaene, Stanislas, *Cum învățăm. De ce este creierul uman mai performant decât orice computer... deocamdată*, trad. Darius Stănescu, Litera, București, 2024.

Dehaene, Stanislas; Lau, Hakwan; Kouider, Sid; „What is consciousness, and could machines have it?”, în *Science*, vol. 358, nr. 6362, 2017, pp. 43–56.

Deng, Li, Yu, Dong, „Deep Learning: Methods and Applications”, în *Foundations and Trends in Signal Processing*, vol. 7, nr. 3–4, 2013, 197–387.

Dennett, Daniel C., „Evolution, Error and Intentionality”, în Y. Wilks and D. Partridge (eds.), *Sourcebook on the Foundations of Artificial Intelligence*, New Mexico University Press, 1988.

Dennett, Daniel C., „Quining Qualia” în A. Marcel, E. Bisiach (eds.), *Consciousness in Modern Science*, Oxford University Press, 1988.

Dennett, Daniel C., *Consciousness Explained*, Back Bay Books, New York, 1991.

Dennett, Daniel C., *From Bacteria to Bach and Back*, W. W. Norton Company, New York, 2017.

Descartes, René, *Discurs asupra metodei*, Editura Științifică, București, 1957.

Dike, Happiness U.; Zhou, Yimin; Deveerasetty, Kranthi K.; Wu, Qingtian; „Unsupervised Learning Based On Artificial Neural Network: A Review”, în *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, China, 2018, pp. 322–327.

Eden, Amnon H.; Steinhart, Eric; Pearce, David; Moor, James H.; „Singularity Hypotheses: An Overview: Introduction to: Singularity Hypotheses: A Scientific and

Philosophical Assessment", în Amnon H. Eden, James H. Moor, Johnny H. Søraker, Eric Steinhart (eds.) *Singularity Hypotheses*, The Frontiers Collection, Springer, Berlin, 2012, pp. 1–12.

Elkins, Katherine; Chun, Jon; „Can GPT-3 Pass a Writer's Turing Test?", în *Journal of Cultural Analytics*, vol. 5, nr. 2, 2020.

Fallis, Don, „What Is Disinformation?", în *Library Trends*, vol. 63, nr. 3, 2015, pp. 401–426.

Ferrara, Emilio; Varol, Onur; Davis, Clayton; Menczer, Filippo; Flammini, Alessandro; „The rise of social bots", în *Communications of the ACM*, vol. 59, nr. 7, 2016, pp. 96–104.

Fetzer, James H., „What is Artificial Intelligence?", în J. Fetzer (ed.), *Artificial Intelligence: Its Scope and Limits*, vol 4, Springer, Dordrecht, 1990.

Fisher, Jillian; Feng, Shangbin; Aron, Robert; Richardson, Thomas; Choi, Yejin; Fisher, Daniel W.; Pan, Jennifer; Tsvetkov, Yulia; Reinecke, Katharina; „Biased LLMs can Influence Political Decision-Making", în *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2025, pp. 6559–6607.

Floridi, Luciano, „The Informational Nature of Personal Identity”, în *Minds & Machines*, vol. 21, nr. 4, 2011, pp. 549–566.

Fodor, Jerry, *The Mind Doesn't Work That Way*, The MIT Press, Cambridge, 2001.

Forlini, Emily, „ Ray Kurzweil: AI Is Not Going to Kill You, But Ignoring It Might”, în *PcMag*, iunie 2024, https://me.pcmag.com/en/ai/23928/ray-kurzweil-ai-is-not-going-to-kill-you-but-ignoring-it-might, accesat la 27.07.2025

Fotopoulou, Aikaterini; Tsakiris, Manos, „Mentalizing homeostasis: the social origins of interoceptive inference – replies to Commentaries”, în *Neuropsychoanalysis*, vol. 19, nr. 1, 2017, pp. 71–76.

Frankish, Keith, „Illusionism”, în J. Symes (ed.), *Philosophers on Consciousness: Talking about the Mind*, Bloomsbury Academic, 2022, pp. 89–99.

Friston, Karl, „The free-energy principle: a rough guide to the brain?”, în *Trends in Cognitive Sciences*, vol. 13, nr. 7, 2009, pp. 293–301.

Friston, Karl; Thornton, Christopher; Clark, Andy; „Free-energy minimization and the dark-room problem”, în *Frontiers in Psychology*, vol. 3, 2012.

Frum, David, „The Very Real Threat of Trump's Deepfake", în *The Atlantic*, aprilie 2020, https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/, accesat la 11.06.2025.

Gallow, J. Dmitri, „Instrumental divergence", în *Philosophical Studies*, vol. 182, nr. 7, 2025, pp. 1581–1607.

Gensler, Harry J., *Introduction to Logic*, ediția a doua, Routledge, New York, 2010.

Gerstgrasser, Matthias; Schaeffer, Rylan; Dey, Apratim; Rafailov, Rafael; Pai, Dhruv; Sleight, Henry; Hughes, John; Korbak, Tomasz; Agrawal, Rajashree; Gromov, Andrey; Roberts, Daniel A.; Yang, Diyi; Donoho, David; Koyejo, Sanmi; „Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data", 2024, https://arxiv.org/pdf/2404.01413v2, accesat la 05.07.2025.

Goel, Ashok K., „Looking back, looking ahead: Symbolic versus connectionist AI", în *AI Magazine*, vol. 42, nr. 4, pp. 83–85.

Gottlieb, Jacqueline; Oudeye, Pierre-Yves; Lopes, Manuel; Baranes, Adrien; „Information-seeking, curiosity, and attention: computational and neural mechanisms" în

Trends in Cognitive Sciences, vol. 17, nr. 11, 2013, pp. 585–593.

Gross, James J., „Emotion regulation: Conceptual and empirical foundations" în J. J. Gross (ed.), *Handbook of emotion regulation*, ediția a doua, The Guilford Press, New York, 2014, pp. 3–20.

Grynbaum, Michael M.; Mac, Ryan; „The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work", în *The New York Times*, decembrie 2023, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html, accesat la 06.09.2025.

Gugerty, Leo, „Newell and Simon's Logic Theorist: Historical Background and Impact on Cognitive Modeling", în *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, nr. 9, 2006, pp. 880–884.

Haidt, Jonathan, *Ipoteza fericirii*, trad. Simona Drelciuc, Humanitas, București, 2020.

Harnad, Stevan, „The symbol grounding problem", în *Physica D: Nonlinear Phenomena*, vol. 42 nr. 1–3, 1990, pp. 335–346.

Haugeland, John, *Artificial Intelligence: The Very Idea*, The MIT Press, Cambridge, 1985.

Hawkins, Jeff, *1000 de creieri: o nouă teorie a inteligenței*, trad. Dan Crăciun, Publica, București, 2022.

Heaven, Will Douglas, „Why Meta's latest large language model survived only three days online", în *MIT Technology Review*, noiembrie 2022, https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/, accesat la 11.06.2025.

Heider, Fritz; Simmel, Marianne; „An Experimental Study of Apparent Behavior", în *The American Journal of Psychology*, vol. 57, nr. 2, 1944, pp. 243–259.

Heritage, Stuart, „'I felt pure, unconditional love': the people who marry their AI chatbots", în *The Guardian*, iulie 2025, https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots, accesat la 01.09.2025.

Hill, Kashmir; Freedman, Dylan; „Chatbots Can Go Into a Delusional Spiral. Here's How It Happens", în *The New York Times*, august 2025, https://www.nytimes.com/2025/08/08/technology/ai-chatbots-delusions-chatgpt.html, accesat la 15.08.2025.

Hochreiter, Sepp, „The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions”, în *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, nr. 2, 1998, pp. 107–116.

Hoffmann, Christian Hugo, „A philosophical view on singularity and strong AI”, în *AI & Society*, vol. 38, 2023, pp. 1697–1714.

Hoffmann, Matej; Pfeifer, Rolf; „The Implications of Embodiment for Behavior and Cognition: Animal and Robotic Case Studies”, https://www.cs.utexas.edu/~jsinapov/teaching/cs378/readings/W7/1202.0440v1.pdf, accesat la 10.06.2025

Hole, Kjell Jørgen; Ahmad, Subutai; „A thousand brains: toward biologically constrained AI”, în *SN Applied Sciences*, vol. 3, nr. 743, 2021.

Hossain Faruk, Md Jobair; Shahriar, Hossain; Valero, Maria; Barsha, Farhat Lamia; Sobhan, Shahriar; Khan, Md Abdullah; Whitman, Michael; Cuzzocrea, Alfredo; Lo, Dan; Rahman, Akond; Wu, Fan; „Malware Detection and Prevention using Artificial Intelligence Techniques”, la *IEEE International Conference on Big Data*, SUA, 2021, pp. 5369–5377.

Howie, David, *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*, Cambridge University Press, Cambridge, 2002.

Hume, David, *An enquiry concerning human understanding*, Oxford University Press, New York, 2007.

James, William, „What is an Emotion?”, în *Mind*, vol. 9, nr. 34, 1884, pp. 188–205.

Jargon, Julie; Kessler, Sam; „A Troubled Man, His Chatbot and a Murder-Suicide in Old Greenwich”, în *The Wall Street Journal*, august 2025, https://www.wsj.com/tech/ai/chatgpt-ai-stein-erik-soelberg-murder-suicide-6b67dbfb, accesat la 01.09.2025.

Jeffrey, Richard C., *Formal Logic: Its Scope and Limits*, ediția a doua, McGraw-Hill, New York, 1981.

Jiang, Bowen; Xie, Yangxinyu; Hao, Zhuoqun; Wang, Xiaomeng; Mallick, Tanwi; Su, Weijie J.; Taylor, Camillo J.; Roth, Dan; „A Peek into Token Bias: Large Language Models Are Not Yet Genuine Reasoners”, în *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 4722–4756.

Jones, Cameron R.; Bergen, Benjamin K.; „Large Language Models Pass the Turing Test”, 2025, https://arxiv.org/abs/2503.23674, accesat la 05.04.2025.

Kahneman, Daniel, *Gândire rapidă, gândire lentă*, trad. Dan Crăciun, Publica, București, 2012.

Kerr, Dara, „OpenAI eyes world's largest valuation for private company in stock sale talks", în *The Guardian*, august 2025, https://www.theguardian.com/technology/2025/aug/19/openai-chatgpt-stock-sale-reports, accesat la 21.08.2025.

Kosmyna, Nataliya; Hauptmann, Eugene; Yuan, Ye Tong; Situ, Jessica; Liao, Xian-Hao; Beresnitzky, Ashly Vivian; Braunstein, Iris; Maes, Pattie; „Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task", 2025, https://arxiv.org/pdf/2506.08872v1, accesat la 01.09.2025.

Kuehn, Johannes; Haddadin, Sami; „An Artificial Robot Nervous System To Teach Robots How To Feel Pain And Reflexively React To Potentially Damaging Contacts", în *IEEE Robotics and Automation Letters*, vol. 2, nr. 1, pp. 72–79.

Kuhn, Thomas, *Structura revoluțiilor științifice*, trad. Radu J. Bogdan, Humanitas, București, 2008.

Lake, Brenden M.; Ullman, Tomer D.; Tenenbaum, Joshua B.; Gershman, Samuel J.; „Building machines that learn and think like people", în *Behavioral and Brain Sciences*, vol. 40, 2017.

Larson, Erik J., *Mitul inteligenței artificiale: de ce computerele nu pot gândi la fel ca noi*, trad. Ovidu-Gheorghe Ruța, Polirom, Iași, 2022.

Lau, Jey Han; Cohn, Trevor; Baldwin, Timothy; Brooke, Julian; Hammond, Adam; „Deep-speare: A joint neural model of poetic language, meter and rhyme", în *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers*, pp. 1948–1958.

LeDoux, Joseph, „Coming to terms with fear", în *Proceedings of the National Academy of Sciences*, vol. 111, nr. 8, 2014, pp. 2871–2878.

LeDoux, Joseph, *ANXIOS. Cum ne ajută creierul să înțelegem și să tratăm frica și anxietatea*, trad. Mihaela Marian Mihăilaș, ASCR, Cluj-Napoca, 2018.

Levesque, Hector; Davis, Ernest; Morgenstern, Leora; „The winograd schema challenge", în *Proceedings of the Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012, pp. 552–561.

Lewis, David K., „An Argument for the Identity Theory", în *The Journal of Philosophy*, vol. 3, nr. 1, 1966, pp. 17–25.

Lighthill, James, „Artificial Intelligence: A General Survey”, în *Artificial Intelligence: A Paper Symposium*, 1973.

Lipton, Peter, *Inference to the Best Explanation*, ediția a doua, Routledge, Londra, 2004.

LoBue, Vanessa; Adolph, Karen E.; „Fear in infancy: Lessons from snakes, spiders, heights, and strangers”, în *Developmental Psychology*, vol. 55, nr. 9, 2019, pp. 1889–1907.

Lorenz, Konrad, *Cele opt păcate capitale ale omenirii civilizate*, Humanitas, București, 2017.

Lovelock, James, *Novacene: The Coming Age of Hyperintelligence*, The MIT Press, Cambridge, 2019.

MacLean, Paul, „The triune brain, emotion, and scientific bias”, în F. O. Schmidt (ed.), *The neurosciences. Second study program*, Rockefeller University Press, New York, pp. 336–349.

Mangan, Bruce, „Dennett, Consciousness, and the Sorrows of Functionalism”, în *Consciousness and Cognition*, vol. 2. nr. 1, 1993, pp. 1–17.

Marcus, Gary, „The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence”, 2020, https://arxiv.org/pdf/2002.06177, accesat la 25.08.2025.

McDermott, Drew, „Artificial intelligence meets natural stupidity”, în *ACM SIGART Bulletin*, vol. 57, 1976, pp. 4–9.

McMahon, Liv, „AI system resorts to blackmail if told it will be removed”, în *BBC*, mai 2025, https://www.bbc.com/news/articles/cpqeng9d20go, accesat la 02.06.2025.

Melnyk, Andrew, „Searle's Abstract Argument against Strong AI”, în *Synthese*, vol. 108, nr. 3, *Computation, Cognition and AI*, 1996, pp. 391–419.

Merker, Bjorn, „Consciousness without a cerebral cortex: a challenge for neuroscience and medicine”, în *Behavioral and Brain Sciences*, vol. 30, 2007, pp. 63–68.

Meyniel, Florent; Dehaene, Stanislas; „Brain networks for confidence weighting and hierarchical inference during probabilistic learning”, în *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, 2017, pp. 3859–3868.

Milmo, Dan, „Man develops rare condition after ChatGPT query over stopping eating salt”, în *The Guardian*, august 2025, https://www.theguardian.com/technology/2025/aug/12/us-man-bromism-salt-diet-chatgpt-openai-health-information, 2025, accesat la 01.09.2025.

Milmo, Dan; Kerr, Dara; „‘It’s missing something’: AGI, superintelligence and a race for the future”, în *The Guardian*, august 2025, https://www.theguardian.com/technology/2025/aug/09/its-missing-something-agi-superintelligence-and-a-race-for-the-future, accesat la 02.09.2025.

Mirzadeh, Iman; Alizadeh, Keivan; Shahrokhi, Hooman; Tuzel, Oncel; Bengio, Samy; Farajtabar, Mehrdad; „GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models”, 2024, https://arxiv.org/pdf/2410.05229, accesat la accesat la 14.06.2025.

Mitchell, Tom Michael, *Machine Learning*, McGraw-Hill Education, New York, 1997.

Miwa, Hiroyasu; Okuchi, Tetsuya; Itoh, Kazuko; Takanobu, Hideaki; Takanishi, Atsuo; „A new mental model for humanoid robots for human friendly communication - introduction of learning system, mood vector and second order equations of emotion”, în *Proceeding of the 2003 IEEE International Conference on Robotics & Automation*, 2003, pp. 3588–3593.

Nagel, Thomas, „What Is It Like to Be a Bat?”, în *The Philosophical Review*, vol. 83, nr. 4, 1974, pp. 435–450.

Nellis, Stephen, „Nvidia CEO says AI could pass human tests in five years”, în *Reuters*, martie 2024, https://www.reuters.com/technology/nvidia-ceo-says-ai-could-pass-human-tests-five-years-2024-03-01/, accesat la 28.06.2025.

Newman, James; Baars, Bernard, „A Neural Attentional Model for Access to Consciousness: A Global Workspace Perspective”, în *Concepts in Neuroscience*, vol. 4, nr. 2, 1993.

Nielsen, Michael A., *Neural Networks and Deep Learning*, Determination Press, 2015.

O’Brien, Matt, „Warner Bros. sues Midjourney for AI-generated images of Superman, Bugs Bunny and other characters”, în *The Associated Press*, septembrie 2025, https://apnews.com/article/warner-bros-midjourney-ai-copyright-lawsuit-dc-studios-b87d80d7b4a4dfdcf0ee149d30830551, accesat la 07.09.2025.

OpenAI: Berner, Christopher; Brockman, Greg; Chan, Brooke; Cheung, Vicki; Dębiak, Przemysław; Dennison, Christy; Farhi, David; Fischer, Quirin; Hashme, Shariq; Hesse, Chris; Józefowicz, Rafal; Gray, Scott; Olsson, Catherine; Pachocki, Jakub; Petrov, Michael; Pondé de Oliveira Pinto, Henrique; Raiman, Jonathan; Salimans, Tim; Schlatter, Jeremy; Schneider, Jonas; Sidor,

Szymon; Sutskever, Ilya; Tang, Jie; Wolski, Filip; Zhang, Susan; „Dota 2 with Large Scale Deep Reinforcement Learning”, 2019, https://cdn.openai.com/dota-2.pdf, accesat la 11.02.2025.

Panksepp, Jaak, *Affective neuroscience: The foundations of human and animal emotions*, Oxford University Press, New York, 1998.

Parrington, John, *Conștiința: cum dă naștere materia gândirii*, trad. Alexandru Babeș, Humanitas, București, 2024.

Peirce, Charles, *Collected Papers of Charles Sanders Peirce*, C. Hartshorne, P. Weiss, A. Burks (eds.), Harvard University Press, Cambridge, 1931–1958.

Perry, R. Michael, „Consciousness as Computation: a Defense of Strong AI Based on Quantum-State Functionalism”, în Charles Tandy (ed.), *Death and Anti-Death, Volume 4: Twenty Years After De Beauvoir, Thirty Years After Heidegger*, Ria University Press, Palo Alto, 2006.

Pillay, Tharin, „How OpenAI’s Sam Altman Is Thinking About AGI and Superintelligence in 2025”, în *TIME*, ianuarie 2025, https://time.com/7205596/sam-altman-superintelligence-agi/, accesat la 10.06.2025.

Placani, Adriana, „Anthropomorphism in AI: hype and fallacy”, în *AI Ethics*, vol. 4, 2024, pp. 691–698.

Polger, Thomas W., „Functionalism as a philosophical theory of the cognitive sciences”, în *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 3, 2012, pp. 337–348.

Puccetti, Roland, „The Great C-Fiber Myth: A Critical Note”, în *Philosophy of Science*, vol. 44, nr. 2, 1977, pp. 303–305.

Putnam, Hilary, „Brains and Behavior”, în R. J. Butler (ed.), *Analytical Philosophy, Second Series*, Basil Blackwell, Oxford, p. 211–235.

Putnam, Hilary, „The Nature of Mental States”, în *Philosophical Papers*, vol. 2, Cambridge University Press, Cambridge, 1975, pp. 429–440.

Raile, Paolo, „The usefulness of ChatGPT for psychotherapists and patients”, în *Humanities and Social Sciences Communications*, vol. 11, 2024.

Romanishyn, Alexander; Malytska, Olena; Goncharuk, Vitaliy; „AI-driven disinformation: policy recommendations for democratic resilience”, în *Frontiers in Artificial Intelligence*, vol. 8, 2025.

Rombach, Robin; Blattmann, Andreas; Lorenz, Dominik; Esser, Patrick; Ommer, Björn; „High-Resolution

Image Synthesis with Latent Diffusion Models", 2021, https://arxiv.org/abs/2112.10752, accesat la 03.07.2025.

Romeijn, Jan-Willem, „Abducted by Bayesians?", în *Journal of Applied Logic*, vol. 11, nr. 4, 2013, pp. 430–439.

Rosenbush, Steven, „Companies Are Struggling to Drive a Return on AI. It Doesn't Have to Be That Way.", în *The Wall Street Journal*, aprilie 2025, https://www.wsj.com/articles/companies-are-struggling-to-drive-a-return-on-ai-it-doesnt-have-to-be-that-way-f3d697aa, accesat la 02.09.2025.

Rosenthal, David, „Higher-order awareness, misrepresentation and function", în *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 367, 2012, pp. 1424–1438.

Russell, Stuart J.; Norvig, Peter; *Artificial intelligence: a modern approach*, Prentice Hall, New Jersey, 1995.

Safiullah, Md.; Parveen, Neha; „Big Data, Artificial Intelligence and Machine Learning: A Paradigm Shift in Election Campaign", în Sandeep Kumar Panda, Ramesh Kumar Mohapatra, Subhrakanta Panda, S. Balamurugan (eds.) *The New Advanced Society: Artificial*

Intelligence and Industrial Internet of Things Paradigm, 2022, pp. 247–261.

Sazlı, Murat H., „A Brief Review of Feed-Forward Neural Networks", în *Communications Faculty of Sciences University of Ankara Series A2-A3 Physical Sciences and Engineering*, vol. 50, nr. 1, 2006, pp. 11–17.

Schaller, R. R., „Moore's law: past, present and future" în *IEEE Spectrum*, vol. 34, nr. 6, 1997, pp. 52–59.

Seager, William, „Consciousness, Value and Functionalism", în *PSYCHE*, vol. 7, nr. 20, 2001.

Searle, John, „Minds, brains, and programs", în *Behavioral and Brain Sciences*, vol. 3, 1980, pp. 417–424.

Searle, John, „Unconsciousness and Intentionality", în *Philosophical Issues*, vol. 1, 1991, pp. 45–66.

Searle, John, *The Rediscovery of the Mind*, The MIT Press, Londra, 1992.

Searle, John, *Mintea: o scurtă introducere în filosofia minții*, trad. Iustina Cojocaru, Herald, București, 2018.

Shannon, Claude, „A Mathematical Theory of Communication", în *The Bell System Technical Journal*, vol. 27, 1948, pp. 379–423.

Shannon, Claude, „Programming a Computer for Playing Chess”, în *Philosophical Magazine*, vol. 41, nr. 314, 1950.

Shao, Chengcheng; Ciampaglia, Giovanni Luca; Varol, Onur; Yang, Kai-Cheng; Flammini, Alessandro; Menczer, Filippo; „The spread of low-credibility content by social bots”, în *Nature Communications*, vol. 9, 2018.

Sheng, Emily; Chang, Kai-Wei; Natarajan, Premkumar; Peng, Nanyun; „The Woman Worked as a Babysitter: On Biases in Language Generation”, în *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3407–3412.

Shoemaker, Sydney, „Some Varieties of Functionalism”, în *Functionalism and the Philosophy of Mind*, vol. 12, nr. 1, 1981, pp. 93–119.

Shumailov, Ilia; Shumaylov, Zakhar; Zhao, Yiren; Papernot, Nicolas; Anderson, Ross; Gal, Yarin; „AI models collapse when trained on recursively generated data”, în *Nature*, vol. 631, 2024, pp. 755–759.

Smart, J. J. C., „Sensations and Brain Processes”, în *The Philosophical Review*, vol. 68, nr. 2, 1959, pp. 141–156.

Smart, J. J. C., „Materialism”, în *The Journal of Philosophy*, vol. 60, nr. 22, 1963, pp. 651–662.

Solms, Mark, *Izvorul ascuns: de ce și cum se naște conștiința*, trad. Dan Bălănescu, Polirom, Iași, 2022.

Solms, Mark; Friston, Karl, „How and Why Consciousness Arises: Some Considerations from Physics and Physiology”, în *Journal of Consciousness Studies*, vol. 25, nr. 5–6, 2018, pp. 202–238.

Sorem, Erik, „Searle, Materialism and the Mind-Body Problem”, în *Perspectives: International Postgraduate Journal of Philosophy*, vol. 3, 2010, pp. 30–54.

Sperry, Roger W., „A modified concept of consciousness”, în *Psychological Review*, vol. 76, nr. 6, 1969, pp. 532–536.

Sriramanan, Gaurang; Bharti, Siddhant; Sadasivan, Vinu Sankar; Saha, Shoumik; Kattakinda, Priyatham; Feizi, Soheil; „LLM-Check: Investigating Detection of Hallucinations in Large Language Models”, în A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (eds.), *Advances in Neural Information Processing Systems*, vol. 37, 2024.

Stephan, Achim, „Varieties of Emergentism”, în *Evolution and Cognition*, vol. 5, nr. 1, 1999, pp. 49–59.

Stove, David, „Hume, Probability, and Induction”, în *The Philosophical Review*, vol. 74 nr. 2, 1965, pp. 160–177.

Sutton, Richard S.; Barto, Andrew G.; *Reinforcement Learning: An Introduction*, ediția a doua, The MIT Press, Cambridge, 2018.

Syed, Shoeb Ali, „AI-Powered Cybercrime: The New Frontier of Digital Threats”, în *International Journal of Engineering Technology Research & Management*, vol. 6, nr. 2, 2022, pp. 214–220.

Thanh, Cong Truong; Zelinka, Ivan; „A Survey on Artificial Intelligence in Malware as Next-Generation Threats”, în *MENDEL*, vol. 25, nr. 2, 2019, pp. 27–34.

Thompson, Brian; Dhaliwal, Mehak Preet; Frisch, Peter; Domhan, Tobias; Federico, Marcello; „A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism”, 2024, https://arxiv.org/pdf/2401.05749, accesat la 3.07.2025.

Trafton, Anne, „Artificial intelligence yields new antibiotic”, în *MIT News*, februarie 2020, https://news.mit.edu/2020/artificial-intelligence-identifies-new-antibiotic-0220, accesat la 02.09.2025.

Turing, Alan M., „Computing Machinery and Intelligence”, în *Mind*, vol. 59, nr. 236, 1950, pp. 433–460.

Turner, Jonathan, „The Evolution of Human Emotions”, în J. Stets, J. Turner (eds.) *Handbook of the Sociology of Emotions*, vol. II, Springer, Dordrecht, 2014, pp. 11–31.

Tversky, Amos; Kahneman, Daniel; „Judgment under Uncertainty: Heuristics and Biases”, în *Science*, vol. 185, nr. 4157, 1974, pp. 1124–113.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz; Polosukhin, Illia; „Attention Is All You Need”, în *Proceedings of the 31st Conference on Neural Information Processing Systems*, SUA, 2017.

Vinge, Vernor, „The coming technological singularity: how to survive in the post-human era”, în *Vision 21: interdisciplinary science and engineering in the era of cyberspace*, NASA: Lewis Research Center, 1993, pp. 11–22.

Von Der Assen, Jan; Celdrán, Alberto Huertas; Luechinger, Janik; Sánchez, Pedro Miguel Sánchez; Bovet, Gérôme; Pérez, Gregorio Martínez; Stiller, Burkhard; „RansomAI: AI-powered Ransomware for Stealthy Encryption”, 2023, https://arxiv.org/pdf/2306.15559, accesat la 02.07.2025.

Wan, Yixin; Pu, George; Sun, Jiao; Garimella, Aparna; Chang, Kai-Wei; Peng, Nanyun; „"Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters", în *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 3730–3748.

Watson, John, „Psychology as the behaviorist views it", în *Psychological Review*, vol. 20, nr. 2, 1913, pp. 158–177.

White, Leslie A., „The Concept of Culture", în *American Anthropologist, New Series*, vol. 61, nr. 2 1959, pp. 227–251.

Whitehead, Alfred North, *Process and Reality*, The Free Press, New York, 1978.

Wiener, Norbert, *The Human Use of Human Beings; Cybernetics and Society*, Free Association Books, Londra, 1989.

Wiener, Norbert, *Invention: The Care and Feeding of Ideas*, The MIT Press, Cambridge, 1994.

Wittgenstein, Ludwig, *Philosophical investigations*, ediția a 3-a, Basil Blackwell, Oxford, 1967.

Yampolskiy, Roman, „Detecting Qualia in Natural and Artificial Agents", 2017, https://arxiv.org/abs/1712.04020, accesat la 04.05.2025.

Yudkowsky, Eliezer, „Artificial Intelligence as a Positive and Negative Factor in Global Risk”, în Eliezer Yudkowsky (ed.), *Global Catastrophic Risks*, Oxford University Press, 2008.

Zahra, Beenish, „Artificial Intelligence and Cyc”, în *Lahore Garrison University Research Journal of Computer Science and Information Technology*, vol. 1, nr. 4, 2017, pp. 29–36.

Zeman, Adam, „Consciousness”, în *Brain*, vol. 124, nr. 7, 2001, pp. 1263–1289.

Webografie

Aschenbrenner, Leopold, https://situational-awareness.ai/, accesat la 29.06.2025.

The Authors Guild, „The Authors Guild, John Grisham, Jodi Picoult, David Baldacci, George R.R. Martin, and 13 Other Authors File Class-Action Suit Against OpenAI”, 2023, https://authorsguild.org/news/ag-and-authors-file-class-action-suit-against-openai/, accesat la 06.09.2025.

Barrett, Lisa Feldman; Solms, Mark; „Discussion between Lisa Feldman Barrett and Mark Solms on the nature of emotion (Part 1)”, https://youtu.be/9yEPHUKyBOM?si=7Ge95tj_nw5RrBqQ, accesat la 23.03.2024.

Buckner, Cameron; Garson, James; „Connectionism”, în Edward N. Zalta, Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, 2024, https://plato.stanford.edu/archives/spr2025/entries/connectionism/, accesat la 25.04.2025.

Camera Deputaților, Urmărirea procesului legislativ PL-x 471/2023, https://www.cdep.ro/pls/proiecte/upl_pck2015.proiect?idp=20853, accesat la 11.06.2025.

DeepMind, „AlphaGo”, https://deepmind.google/research/breakthroughs/alphago/, accesat la 14.02.2025.

Douven, Igor, „Abduction”, în Edward N. Zalta, Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, 2021, https://plato.stanford.edu/archives/spr2025/entries/abduction/, accesat la 27.04.2025.

Levin, Janet, „Functionalism”, în Edward N. Zalta, Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, 2023, https://plato.stanford.edu/archives/sum2023/entries/functionalism/, accesat la 10.02.2025.

Marcus, Gary, „GPT-5: Overdue, overhyped and underwhelming. And that’s not the worst of it.”, https://garymarcus.substack.com/p/gpt-5-overdue-overhyped-and-underwhelming, accesat la 02.09.2025.

Marcus, Gary, „Humans versus Machines: The Hallucination Edition”, https://garymarcus.substack.com/p/humans-versus-machines-the-hallucination, accesat la 15.07.2025.

Marcus, Gary, „What to Expect When You’re Expecting... GPT-5”,

https://garymarcus.substack.com/p/what-to-expect-when-youre-expecting-62e, accesat la 02.08.2025.

Nagel, Thomas, „Psychophysical Monism as an Ideal", *the 26th meeting of the Association for the Scientific Study of Consciousness*, New York University, 2023, https://youtu.be/VU4u-LfkI7w?si=kuPHxmQJrDCq3g2t, accesat la 28.02.2024.

OpenAI, https://openai.com/index/building-more-helpful-chatgpt-experiences-for-everyone/, accesat la 02.09.2025.

Quantum Techscribe, „A Comprehensive Guide to the History of Artificial Intelligence: Foundational Concepts and Early Developments", 2025, https://quantumzeitgeist.com/history-of-artificial-intelligence/, accesat la 02.09.2025.

Robinson, Howard, „Dualism", în Edward N. Zalta, Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*, 2023, https://plato.stanford.edu/archives/spr2023/entries/dualism/, accesat la 21.02.2025.