GAURAV SURI JAY McCLELLAND

# Mintea emergentă

## Despre cum apare inteligența la oameni și mașini

Traducere din engleză de
Adina Avramescu

# Cuprins

# Prefață

Creierul nostru este o rețea vastă de celule numite *neuroni,* fiind animată de tipare de activitate electrică și chimică care crește, scade și crește din nou. Percepțiile, gândurile, deciziile și acțiunile noastre – procesele pe care le vom numi *mintea* – apar din aceste tipare de activitate.

Dar *cum*? Cum e posibil ca mintea să ia naștere din tiparele de activitate ale creierului?

Pentru noi, aceasta reprezintă una dintre cele mai persistente și mai tulburătoare întrebări ale omenirii. Ea privește însăși natura noastră și locul nostru în univers. Și vizează de asemenea posibilitatea existenței – și, în caz că există, natura – minților artificiale. Noi am scris *Mintea emergentă* pentru a împărtăși un nou tip de răspuns la această întrebare.

Suntem oameni de știință practicieni care ne-am dedicat carierele înțelegerii minții umane. Când am început călătoriile noastre inițiale separate, fiecare am căutat să aflăm dacă funcționarea minții poate fi înțeleasă într-o manieră *mecanică* – în felul în care cineva ar putea încerca să înțeleagă, de pildă, un avion sau modul în care un virus provoacă o boală. Am descoperit că abordările existente legate de întrebările noastre erau vag formulate și adesea nu făceau trimitere la fapte relevante despre activitatea cerebrală. Ne-am gândit că, dacă ne întemeiem explicațiile pe modul mecanic de funcționare a creierului, acest lucru ne-ar putea oferi răspunsuri mai bune la întrebările despre noi înșine.

Când a murit, fizicianul Richard Feynman a lăsat scris pe tabla lui: „Ce nu pot să creez, nu înțeleg". Afirmația lui surprinde un aspect esențial al abordării noastre. Noi ne-am propus să construim sisteme asemănătoare creierului care recreează fenomene ale minții pe care căutăm să le înțelegem. Dar recrearea unui creier în toate detaliile sale este un lucru nerealist. Prin urmare, pentru a ne ajuta să evoluăm, noi construim modele ce elimină multe detalii.

Modelele pe care le folosim – numite *modele de rețele neuronale* – ne-au fost inspirate de vastele rețele de neuroni din creier. Ele ne permit să explorăm modul în care capacitățile noastre umane ar putea apărea din activitatea neuronală. Aceste modele ignoră în mod deliberat multe dintre complexitățile creierului pentru a se concentra pe procesele elementare care ne ajută să înțelegem cum funcționează mintea noastră. În această carte vom descrie rețelele neuronale care contribuie la clarificarea modului în care oamenii percep, decid, formează concepte și urmăresc obiective.

Astfel de modele ne ajută de asemenea să lămurim răspunsuri la întrebări despre noi, oamenii, care ne-au nedumerit și pe noi, și pe mulți alții. Aceste întrebări încep adesea cu cuvinte precum *de ce*, *ce* și *unde*. De ce nu reușim uneori să ne transformăm intențiile în fapte? De ce manifestăm, noi și alții, prejudecăți adânc înrădăcinate? Ce anume din natura noastră ne permite uneori să vedem cu ușurință că ceva este adevărat, în vreme ce alteori nu reușim să înțelegem? De unde vin intuițiile noastre și de ce pot ele să fie adesea greșite?

Un lucru remarcabil este că rețelele neuronale – transpuse în programe de calculator – au devenit fundamentul inteligenței artificiale. Modelele pe care noi și alții le-am dezvoltat inițial pentru a înțelege mintea umană s-au dovedit a fi baza pentru construirea minților artificiale. Astfel, înțelegerea modului în

care rețelele neuronale surprind capacitățile noastre asemănătoare gândirii aruncă, de asemenea, o lumină asupra sistemelor IA din ziua de azi. În această carte discutăm principiile-cheie ale modelelor rețelelor neuronale ale propriilor minți care stau la baza sistemelor IA ce ating și uneori depășesc anumite aspecte ale abilităților noastre cognitive umane. Deși ne concentrăm pe ceea ce noi considerăm a fi idei de durată, discutăm selectiv și anumite inovații care au evoluat rapid în sistemele IA începând din 2020 – în special în ultimele două capitole. Aceste inovații, interesante și generatoare, s-ar putea să evolueze și mai mult în anii care vin. De câte ori a fost posibil, noi am descris asemenea idei într-un fel care scoate în evidență mai degrabă principiile care stau la baza lor, decât implementările lor tranzitorii. Scopul nostru este să oferim cititorilor un fundament care rămâne util chiar și atunci când implementările specifice evoluează.

Cartea este structurată în patru părți. În Partea întâi începem prin a descrie modul în care un sistem poate avea proprietăți care nu sunt prezente în niciuna dintre componentele sale. Acest fenomen – numit *emergență* – este central pentru cadrul teoretic al rețelelor neuronale privind mintea, conform căruia mintea emerge din interacțiunile dintre unități de procesare simple, asemănătoare celulelor cerebrale. Aceste celule, luate individual, nu pot gândi, dar interacțiunea lor activează un sistem care gândește. În Partea a doua prezentăm modul în care cadrul teoretic al rețelelor neuronale privind mintea poate contribui la explicarea unei game variate de comportamente umane. Mai întâi, examinăm explicațiile bazate pe rețele neuronale ale memoriei – inclusiv ale caracterului ei failibil. Apoi analizăm dependența noastră de context în ce privește înțelegerea lumii din jurul nostru – inclusiv modul în care așteptările noastre ne modelează gândurile. În

continuare, examinăm procesul nostru decizional – inclusiv modul în care alegerile noastre sunt uneori raționale, iar alteori irațională. În Partea a treia detaliem modul în care rețelele neuronale – atât cele biologice, cât și cele artificiale – învață din experiență. Descriem cum învățarea face posibilă cunoașterea obiectelor și a proprietățile lor și cum poate ea susține folosirea limbajului în special în modele lingvistice de mari dimensiuni (*large language models* – LLM-uri). În cele din urmă, în Partea a patra extindem și aplicăm perspectiva rețelelor neuronale asupra minții. Descriem modul în care rețelele neuronale pot fi utile pentru înțelegerea unor fenomene precum raționamentul formal, comportamentul motivat și conștiința – aspecte ale minții care nu au fost încă integrate pe deplin în rețelele neuronale. Încheiem cu discutarea unor implicații ale perspectivei rețelei neuronale – atât pentru oameni, cât și pentru mașini.

Pe parcursul cărții am inclus interludii care sunt, dacă nu se precizează altfel, fanteziste și fictive. De exemplu, într-o conversație apare Sigmund Freud vorbind cu Adam Smith, în alta apare editorul acestei cărți vorbind cu un client (fictiv) la un bar din New York. Aceste conversații și toate cuvintele atribuite oamenilor – fie ei reali sau fictivi – sunt inventate de noi, deci plăsmuiri ale imaginației noastre. Sperăm că aceste conversații îți vor anima reflecțiile asupra chestiunilor în discuție la fel de mult pe cât ne-au animat nouă procesul de scriere a acestei cărți.

Am scris această carte pentru oricine este curios să afle mai multe lucruri atât despre mintea umană, cât și despre cea artificială. Nu e nevoie de niciun fel de cunoștințe sofisticate de matematică. În afară de simpla adunare și înmulțire, nu există nicio ecuație în carte. De asemenea, nu e nevoie de niciun fel de cunoștințe anterioare legate de științele cognitive, psiholo-

gice, neuronale sau informatice. La sfârșitul cărții oferim note cu trimiteri la surse pentru cei care sunt interesați să afle mai mult.

Înțelegerea noastră referitoare la mecanismele minții evoluează în permanență. Sunt multe lucruri care au rămas nedescoperite – dar ceea ce se știe deja este impresionant și cu semnificații profunde.

Te invităm să pornești în această călătorie cu noi. Poate că îți va îmbogăți felul în care te vezi pe tine și locul tău în universul în care trăim.

*Gaurav și Jay*

# Partea întâi | *Mintea ta este o rețea neuronală*

În Partea întâi te invităm să reflectezi la ideea că mintea noastră este concepută, într-o manieră utilă, ca apărând din interacțiunea dintre celulele creierului care, luate individual, nu au capacitățile minții. Noi introducem ideea de emergență – un fenomen în care întregul are proprietăți care nu sunt prezente în niciuna dintre părțile sale – și descriem cum modelele rețelelor neuronale ne ajută să înțelegem emergența minții.

# 1 | O invitație

Când unul dintre autorii acestei cărți, Gaurav, avea 14 ani, părinții i-au dat cadou de ziua lui de naștere echivalentul a 50 de dolari. Banii îi ajungeau să își cumpere o pereche de blugi evazați – care erau la modă printre adolescenți la vremea aceea – sau să își rezerve un loc într-o excursie școlară mult dorită, alături de mulți dintre prietenii săi. Problema era că nu putea să aibă decât unul dintre aceste lucruri, dar și le dorea din tot sufletul pe amândouă. Trebuia să ia o decizie. Așadar, în acea seară, și-a luat inima-n dinți și a ales. Avea să meargă în excursie. La urma urmei, pantalonii mai puteau aștepta. Era convins că făcea ce trebuie.

Dar în dimineața următoare, când s-a trezit, s-a întâmplat ceva neașteptat: Gaurav era sigur că trebuie să aleagă pantalonii. Nu se schimbase nimic referitor la cele două opțiuni, dar acum decizia lui era alta. Această oscilație, care s-a repetat de câteva ori pe parcursul următoarelor câteva zile, l-a nedumerit foarte mult. La acea vreme, tocmai începuse să lucreze pe calculatoare și își imagina că mintea era un fel de calculator care funcționa după principii logice. Ce fel de calculator îți dă un răspuns seara și altul dimineața? Cum e posibil ca răspunsurile bazate pe logică să se schimbe aparent fără niciun motiv? Iar dacă gândurile și preferințele lui nu sunt rezultatul logicii și rațiunii, atunci ce fel de lucruri sunt ele?

Gaurav dăduse peste câteva dintre întrebările constante ale omenirii: cum apar gândurile noastre? De ce facem lucrurile

pe care le facem? Putem avea încredere în ceea ce gândim? Și, la modul mai general, ce este mintea și cum funcționează?

***Concepții comune despre minte (și limitele lor)***

Ce este mintea? Am putea-o vedea ca pe ceva aflat în interiorul nostru care dă naștere la gândurile, percepțiile, amintirile, sentimentele, deciziile și acțiunile noastre. Dar ce este ea, de fapt? De unde vine? În cele ce urmează vom prezenta succint câteva concepții comune si limitele lor.

O primă concepție despre minte se întemeiază pe tradițiile noastre religioase, dintre care multe tind spre ideea că mintea noastră derivă din materia divină sau dintr-un lucru spiritual care poate anima trupul uman, transformându-l în persoană. Nu este dificil să înțelegem sursa acestei idei: trupul uman pare un lucru banal și lipsit de judecată. Cum ar putea să producă inteligență? Evident, mintea trebuie să apară din altceva – din ceva nefamiliar, nepământean. Din ceva care ne conectează cu ceea ce este nemuritor, sacru și plin de sens. Din ceva de natură divină.

Ideea că mintea este atributul a ceva etern este fascinantă și chiar frumoasă, dar nu poate să reprezinte o explicație reală a ceea ce este mintea și a modului în care apar gândurile. Dimpotrivă, ea tratează mintea ca pe o entitate inefabilă care se opune unei înțelegeri mai profunde. Dacă scopul este să înțelegem cum produce mintea gânduri și alte lucruri pe care i le atribuim, acesta nu este un punct final acceptabil.

O a doua concepție sugerează că mintea este un set de convingeri și dorințe asemănătoare propozițiilor. Astfel, „Persoanele cu studii universitare câștigă mai mulți bani" este un exemplu de convingere. Și „Aș vrea să obțin în cele din urmă un job bine plătit" este un exemplu de dorință. Pare logic că dorințele și convingerile pot interacționa pentru a produce

intenții și acțiuni. Dacă cineva ne întreabă de ce am decis să acționăm într-un anumit fel, putem indica convingerile și dorințele care păreau să ne conducă decizia. De ce ne-am înscris la facultate? Am putea răspunde: „Pentru a putea obține un job mai bine plătit". Poate că funcționarea minții implică pur și simplu interacțiuni între convingeri și dorințe în vederea formării unor scopuri care ne pot ghida comportamentul.

Un neajuns al modelului minții de tip convingere-dorință este că nu ne explică de unde vin convingerile și dorințele. Cum pot rezulta asemenea lucruri abstracte precum convingerile și dorințele din procese fizice care au loc în creierul nostru și cum pot ele să dea naștere unor acțiuni fizice, incluzând mișcarea, acțiunea și producerea limbajului? Mai mult decât atât, modelul nu explică de ce, în mod frecvent, oamenii nu acționează potrivit convingerilor și dorințelor lor. De exemplu, adesea, pacienții nu iau medicamente esențiale pentru sănătatea lor, iar angajații nu își deschid conturi de pensie importante pentru viitorul lor financiar. Aceste lucruri se întâmplă chiar dacă asemenea indivizi au convingeri pozitive despre eficacitatea tratamentului lor și își doresc siguranța oferită de conturile de pensie. Cu toate acestea, nu reușesc să acționeze.

Potrivit unei alte concepții, mintea este asemănătoare cu un software care colectează date din lumea externă cărora le aplică un set de reguli, probabil furnizate de evoluție. Într-adevăr, în unele cazuri, funcționarea minții pare să implice aplicări ale regulilor: dacă un animal are aripi și poate zbura, îl vom clasifica probabil drept pasăre; alegem un anumit preparat din meniul unui restaurant pentru că credem că ne maximizează valoarea comparativ cu celelalte preparate disponibile; și anticipăm că forma de trecut a unui verb recent inventat precum *fax* (a trimite un fax) este *faxed* (a trimis un fax), în conformitate cu regula ușor de enunțat potrivit căreia trecutul

oricărei părți de vorbire clasificate drept verb se formează prin adăugarea terminației *ed*.

Problema cu perspectiva asupra minții văzute ca software este că nu ne poate duce prea departe în înțelegerea ei. Da, multe păsări pot zbura, dar noi putem recunoaște o pasăre care nu poate zbura. Da, noi alegem adesea din meniu preparatele pe care le preferăm, dar alegerile noastre reflectă adesea influența unor variabile care sunt incidentale în raport cu valoarea inerentă – cum ar fi faptul că un fel de mâncare se află în partea de sus a paginii. Și, da, adesea adăugăm un *ed* pentru a exprima timpul trecut, dar există multe verbe neregulate cărora nu li se poate aplica această regulă (*sleep*, de exemplu, devine *slept*). O altă limită majoră a acestor modele este că ele nu au condus la sisteme IA funcționale. Abordarea pe care o prezentăm în *Mintea emergentă* s-a dovedit mult mai reușită.

O ultimă concepție pe care o întâlnim frecvent afirmă că diferitele aspecte ale minții depind de specializări diferite, fiecare dintre acestea fiind situată în propria regiune specializată a creierului. Potrivit acestei perspective, ne mișcăm datorită regiunilor creierului specializate în mișcare, vedem datorită regiunilor creierului specializate în vedere, ajungem să fim motivați datorită regiunilor creierului specializate în motivație și vorbim datorită regiunilor creierului specializate în limbaj – și, în unele versiuni ale unor asemenea teorii, gândim folosind regiuni specializate care gândesc.

Este de netăgăduit că unele regiuni ale creierului manifestă un oarecare grad de specializare. Întrebarea este ce dă naștere unei asemenea specializări. O abordare, susținută de filosoful Jerry Fodor în cartea sa *The Modularity of Mind*, este că această specializare apare datorită proprietăților interne specifice ale acelor regiuni ale creierului, selectate de evoluție pentru a

efectua calcule specializate pentru sarcinile pe care le îndeplinesc. Deși regiunile creierului diferă între ele într-o anumită măsură în plan intern, perspectiva pe care o prezentăm în *Mintea emergentă* este că o asemenea specializare este în mare parte o consecință a diferențelor dintre conexiunile de intrare și ieșire ale diferitelor regiuni ale creierului. De exemplu, acea parte a creierului numită *cortex vizual* joacă un rol important în percepția vizuală, deoarece primește informații deosebit de puternice de la ochi. Acest lucru sugerează că modificarea stimulilor care ajung într-o regiune a creierului ar trebui să producă o modificare la nivelul funcției îndeplinite de acea regiune a creierului. Într-adevăr, persoanele al căror cortex vizual nu primește stimuli vizuali pentru că sunt oarbe din naștere reutilizează acea parte a creierului pentru sarcini nonvizuale, precum procesarea stimulilor auditivi sau tactili, care furnizează, ambele, anumite informații acestei regiuni a creierului.

În loc să vadă mintea ca pe o colecție de module strict specializate, această perspectivă ne invită să o considerăm mai degrabă ca pe un sistem adaptabil modelat de experiență, de învățare și de cerințele mediului. Această abordare ne ajută să explicăm existența a ceea ce unii oameni de știință numesc „zona formei vizuale a cuvintelor" – o regiune a creierului care pare a fi specializată pentru citirea cuvintelor prezentate vizual, tipărite. Nu este plauzibil să sugerăm că evoluția a selectat această regiune pentru a se specializa în citit, deoarece scrisul și cititul au fost inventate cu circa 5 000 de ani în urmă. Este o perioadă mult prea scurtă pentru ca evoluția să fi putut să producă un modul de citit specializat prin selecție naturală.

Și totuși, această regiune cerebrală se specializează în citit pentru oamenii care au învățat să citească. De ce? Cititul