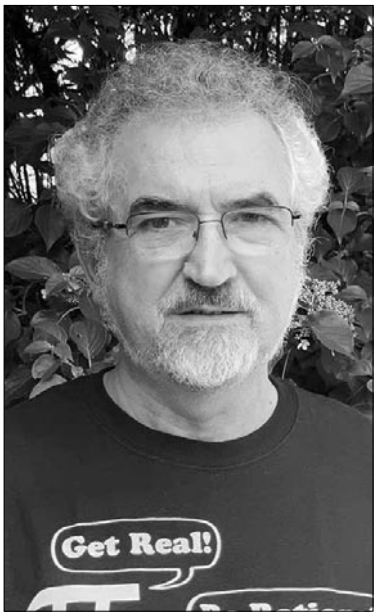
LITERATURA GENERATĂ DE INTELIGENȚA ARTIFICIALĂ sau CE TREBUIE SĂ ȘTIE UN SCRIITOR DESPRE MAȘINA CARE «SCRIE»



Un interviu cu IOAN ROXIN,

profesor emerit la Universitatea Marie et Louis Pasteur (Franţa), fost director al laboratorului de cercetare interdisciplinar ELLIADD, autor al cărţii „Au cœur de l'IA générative. Fonctionnement, espace latent et illusions cognitives", care urmează să apară la FYP Éditions, în iunie 2026.*

În luna februarie, când am iniţiat acest dialog, mi-aţi scris:

„Sunt un om care lucrează de peste patruzeci de ani în informatică, în tehnologia informaţiei, în cercetare universitară, şi care cunoaşte din interior mecanismele inteligenţei artificiale generative. De doi ani lucrez la o carte dedicată tocmai acestei probleme — *Au cœur de l'IA générative. Fonctionnement, espace latent et illusions cognitives* — care va apărea la FYP Éditions în iunie 2026 (cred că va fi publicată şi în limba română). Scopul cărţii este exact acela de a explica pentru un public non-tehnic (inclusiv pentru literaţi şi specialişti în ştiinţele sociale) ce se petrece efectiv în interiorul acestor sisteme — un «interior» pe care simpla utilizare a chatboţilor nu îl face vizibil."

Până la apariţia cărţii, pe care firește că abia o aştept, v-am rugat să răspundeţi la câteva întrebări, în speranţa că prin ele cititorii noştri vor putea anticipa substanţa lucrării dumneavoastră.

Preambul

Când un poet folosește un model de limbaj și obține un text care seamănă cu un poem, el vede rezultatul. Ceea ce nu vede — și ceea ce schimbă radical interpretarea rezultatului — este procesul. A evalua literatura generată de IA fără a înțelege cum funcționează generarea este ca și cum ai evalua o pictură fără a ști că este o fotografie retușată: judecata estetică se schimbă fundamental odată ce cunoști natura obiectului.

1. Când un model de limbaj precum ChatGPT sau Claude „scrie" un poem sau un roman, ce se petrece efectiv în interiorul sistemului? Cum funcţionează generarea de text, explicată pentru un public literar?

- Când un scriitor se aşază la masa de lucru, aduce cu sine tot ce a trăit — lecturi, suferinţe, bucurii, obsesii, ritmuri interioare. Când alege un cuvânt în locul altuia, alegerea vine din întreaga sa fiinţă: simte că un cuvânt e just, că altul sună fals, că un al treilea ar fi mai muzical dar mai puţin precis. Scriitorul scrie *aşa* pentru că nu poate scrie *altfel*.

Ce se petrece în interiorul unui model de limbaj (ChatGPT, Claude, DeepSeek, Gemini, Grok, Mistral) este un proces radical diferit, deşi rezultatul poate, la suprafaţă, să pară similar.

Un model de limbaj a fost „antrenat" pe sute de miliarde de cuvinte: literatură, presă, enciclopedii, articole ştiinţifice, conversaţii, eseuri, poezie. Antrenamentul nu înseamnă lectură, nici memorare în sens uman. Înseamnă extragerea şi înregistrarea, sub formă de parametri numerici (miliarde de „ponderi" matematice), a tiparelor statistice ale limbajului: ce cuvinte urmează de obicei după alte cuvinte, ce structuri gramaticale sunt frecvente, ce asocieri metaforice revin. Modelul nu înţelege niciun text — dar internalizează, cu o precizie remarcabilă, regularităţile tuturor textelor.

Apoi, în momentul generării, pornind de la textul primit de la utilizator, modelul calculează, cuvânt cu cuvânt, probabilitatea fiecărui cuvânt posibil din vocabularul său. Nu „alege" în sensul uman — eşantionează dintr-o distribuţie de probabilitate. Dacă tocmai a generat „*Şi dacă ramuri bat în…*", calculează că „*geam*" are probabilitatea cea mai mare, iar „*noapte*" sau „*vânt*" au probabilităţi mai mici. Produce un cuvânt, trece la următorul, şi la următorul — fiecare depinzând de toate cele anterioare, fiecare fiind rezultatul unui calcul matricial, nu al unei deliberări. Nu există, în niciun moment, ezitare, insatisfacţie, inspiraţie sau necesitate.

Cea mai apropiată analogie pe care o pot oferi este aceasta: imaginaţi-vă un scriitor care a citit toată literatura lumii — fiecare roman, fiecare poem, fiecare eseu — dar care este orb din naştere. N-a văzut niciodată un apus, n-a atins o faţă, n-a mirosit o floare. Ştie cuvântul „*apus*" şi ştie, din milioane de texte, ce cuvinte urmează după „*apus*". Când scrie „*apusul însângerat se stingea dincolo de dealuri*", propoziţia este impecabilă — dar vine din experienţa *propoziţiilor* despre apusuri, nu din experienţa unui apus.

Acest scriitor orb cunoaşte totul despre limbaj şi nimic despre lume.

Sau, în altă formulare: un model de limbaj este ca un pianist extraordinar care nu-şi aude muzica. Degetele lui ating clapele în ordinea corectă — a învăţat milioane de partituri — dar nu aude sunetul, nu simte emoţia, nu reacţionează la ceea ce produce. Execuţia poate fi impecabilă. Interpretarea este absentă.

Există o a treia analogie, poate cea mai lămuritoare: **copistul medieval**. Imaginaţi-vă un călugăr care copiază manuscrise în latină toată viaţa. Nu ştie latină, dar după zeci de ani poate anticipa ce literă urmează după alta, ce formulă completează un paragraf. Dacă i-ai cere să „scrie" un text latin, ar produce ceva care seamănă cu latina, fără a înţelege ce spune. Antrenamentul unui model de limbaj este echivalentul computaţional al acestei copieri nesfârşite:

miliarde de texte sunt parcurse, tiparele lor sunt înscrise în parametrii reţelei. Modelul nu înţelege textele — le **internalizează statistic**. Ştie că după „*a fi sau*" urmează probabil „*a nu fi*" — nu pentru că a meditat la Hamlet, ci pentru că tiparul a fost suficient de frecvent.

Ce rezultă din acest mecanism? Un sistem care produce text fluent, coerent gramatical, stilistic versatil — capabil să imite orice registru, de la liric la epic, de la ironic la elegiac. Dar un sistem care nu are voce proprie (poate scrie în toate stilurile tocmai pentru că nu are niciunul), nu are necesitate interioară (scrie pentru că primeşte un prompt, nu pentru că nu poate să nu scrie), nu are memorie afectivă (cuvântul „*tei*" este un punct într-un spaţiu matematic, nu mirosul copilăriei), şi nu riscă nimic scriind (riscul presupune o miză — pe care maşina nu o are).

A înţelege acest mecanism nu înseamnă a-l trivializa. Complexitatea computaţională este remarcabilă, iar rezultatele pot fi impresionante. Dar a fi impresionat fără a înţelege mecanismul este exact definiţia iluziei — şi tocmai de aceea, înainte de a judeca dacă maşina „creează" sau nu, trebuie să înţelegem ce face efectiv atunci când pare că scrie.

2. Ce este „spaţiul latent" — acest concept-cheie al inteligenţei artificiale generative — şi de ce este el esenţial pentru a înţelege ce poate şi ce nu poate face IA în domeniul creaţiei literare?

- Imaginaţi-vă o hartă imensă a literaturii. Pe această hartă, fiecare text care a existat vreodată ocupă un loc. Bacovia se află într-o regiune — să-i spunem zona melancoliei urbane. Aproape de el, Baudelaire, Trakl, poate fragmente din Cioran. Mai departe, Rabelais, în zona exuberanţei carnavaleşti. Şi mai departe, romanele poliţiste, manualele de fizică, reţetele culinare — fiecare în zona sa. Textele care se aseamănă sunt aproape unele de altele; textele care diferă sunt departe. Această hartă nu este o metaforă — ea există, sub formă matematică, în interiorul fiecărui model de limbaj. Este spaţiul latent.

În cursul antrenamentului, fiecare cuvânt, fiecare expresie, fiecare fragment de text este transformat într-un vector — un şir de numere care constituie coordonatele sale în acest spaţiu. „*Tristeţe*" şi „*melancolie*" au coordonate apropiate; „*tristeţe*" şi „*şurub*" au coordonate îndepărtate. Dar spaţiul nu se limitează la cuvinte izolate — el înregistrează relaţii, structuri, tipare complexe: naraţiuni, ritmuri, asocieri metaforice, registre stilistice. Întreaga experienţă lingvistică a umanităţii se află acolo, comprimată sub formă de geometrie numerică.

Termenul „latent" vine din latină (*latens* — ascuns) şi este esenţial: acest spaţiu este invizibil. Utilizatorul nu-l vede, nu-l simte, nu bănuieşte existenţa lui. Vede doar textul generat — fără a şti că acel text este traiectoria unei deplasări prin acest spaţiu matematic ascuns. Când modelul generează un poem, el navighează în spaţiul latent, trecând dintr-un punct în altul, urmând relieful creat de miliardele de texte absorbite.

Şi aici intervine consecinţa decisivă.

Modelul poate ajunge în orice punct de pe hartă şi, mai ales, *între* punctele existente. Poate genera un text la jumătatea distanţei dintre Bacovia şi Baudelaire — un hibrid care nu există în niciun corpus, dar care este derivat din texte existente. Aceasta este sursa aparenţei de originalitate: textul pare nou pentru că nu este identic cu niciunul existent, dar este derivat din textele existente, o medie ponderată a vecinătăţii sale în spaţiu.

Limita: modelul nu poate ieşi de pe hartă. Nu poate genera un text într-o zonă unde nu există date de antrenament. Nu poate inventa ceea ce Kafka a inventat — un teritoriu literar care nu exista pe nicio hartă şi pe care Kafka l-a creat prin actul scrierii. Kafka nu a interpolat între texte existente — a **extins harta**. Iar extinderea hărţii, prin definiţie, nu poate fi realizată din interiorul ei.

Pentru a preciza această distincţie, propun o a doua imagine: spaţiul latent ca peisaj montan. Văile sunt zonele dense — acolo unde există multe texte similare: romane de dragoste convenţionale, articole de presă, poezie lirică de serie. Crestele sunt zonele rare — scriitura cu adevărat originală. Modelul este ca o bilă care se rostogoleşte pe acest peisaj: gravitaţia probabilistică o trage mereu spre vale, spre

frecvent, spre convenţional. Ca să ajungă pe o creastă ar trebui să urce, dar nu are energie proprie. De aceea textele generate tind structural spre mediana gustului: nu din cauza unor „preferinţe" conservatoare, ci pentru că arhitectura spaţiului latent favorizează centrul, nu marginile. Or, creaţia autentică se petrece la margini.

Când un poet scrie un vers, acel vers vine dintr-un *loc* — o experienţă de viaţă, o obsesie, o viziune. Când un model generează un vers, acel vers vine dintr-un *spaţiu matematic*. Ambele produc text. Dar primul creează un teritoriu nou pe harta literaturii, în timp ce al doilea traversează teritoriul deja cartografiat. Diferenţa este invizibilă la suprafaţa textului — dar este totală la nivelul procesului. Şi tocmai de aceea spaţiul latent este conceptul-cheie: fără el, confundăm traversarea cu descoperirea, interpolarea cu creaţia, navigaţia cu explorarea.

Într-o formulare condensată: IA generativă explorează cu o eficienţă fără precedent teritoriul cunoscut al limbajului. Dar nu poate descoperi continente noi. Descoperirea presupune o navă care pleacă *de pe* hartă — şi IA nu are navă, pentru că IA *este* harta.

3. Ce „iluzii cognitive" ne fac să confundăm un text generat de IA cu o creaţie literară autentică? De ce creierul nostru este „păcălit" atât de uşor?

- Să facem un experiment. Citiţi următorul fragment:

„Seara cade peste grădină ca o mână care acoperă o rană. Teii au obosit de propria lor umbră. Undeva, departe, un câine latră în numele tuturor câinilor care au lătrat vreodată şi nimeni nu i-a răspuns."

Dacă aţi simţit ceva — o undă de melancolie, un fior estetic —, reţineţi senzaţia. Vom reveni la ea.

Deşi acum ştim cum funcţionează generarea de text şi unde se petrece, o întrebare tulburătoare rămâne: de ce nu vedem diferenţa dintre un text generat şi o creaţie autentică?

Răspunsul nu ţine de calitatea textelor generate. Ţine de arhitectura creierului care le citeşte. Suntem victimele unor iluzii cognitive sistematice, la fel de puternice ca iluziile optice — şi la fel de inevitabile.

Fenomenul nu e nou. În 1966, informaticianul Joseph Weizenbaum a creat ELIZA — un program rudimentar care simula un psihoterapeut, reformulând sub formă de întrebare ce-i spunea utilizatorul. „Sunt trist." „De ce spui că eşti trist?" Programul nu înţelegea absolut nimic — aplica reguli mecanice de substituţie. Şi totuşi, utilizatorii se ataşau de ELIZA, îi atribuiau empatie, refuzau să creadă că e doar un program. Weizenbaum a fost atât de tulburat de această reacţie încât a scris o carte întreagă (Computer Power and Human Reason, 1976) pentru a trage un semnal de alarmă. Ceea ce s-a numit de atunci „efectul ELIZA" — tendinţa de a atribui înţelegere unui sistem care nu posedă niciuna — este azi amplificat de milioane de ori. Dacă un program de câteva sute de linii producea ataşament în 1966, imaginaţi-vă ce produc modelele actuale, cu miliardele lor de parametri şi fluenţa lor impecabilă.

Prima iluzie se numeşte euristica fluenţei: cu cât un mesaj e mai uşor de procesat, cu atât creierul îl evaluează automat ca fiind mai adevărat, mai valoros. Or, modelele de limbaj sunt, prin construcţie, optimizate pentru fluenţă — întregul proces de ajustare constă în a face textul cât mai plăcut, cât mai clar. Textele IA primesc astfel un bonus cognitiv pe care nu-l merită. „*Scrie bine, deci gândeşte*" — iată capcana. Paradoxul e că un profund gânditor poate scrie stângaci — Heidegger, de exemplu —, iar cel mai gol mecanism poate scrie impecabil. Confundăm ambalajul cu conţinutul.

Dar fluenţa singură nu explică totul. Reveniţi la fragmentul de la început. Aţi simţit ceva — este normal. Creierul activează automat mecanismele de empatie în prezenţa cuvintelor asociate emoţiei. Când citim „*rană*", „*obosit*", „*nimeni nu i-a răspuns*", circuitele empatice se declanşează ca şi cum am fi în prezenţa cuiva care suferă. Ne emoţionăm, apoi atribuim emoţia sursei: „*textul acesta e sensibil*". Dar textul nu e sensibil — *cititorul* e sensibil. „*Mă emoţionează, deci simte*" — a doua capcană. Emoţia e ca plânsul la un mormânt gol: lacrimile sunt reale, dar mormântul e gol.

Fragmentul de mai sus a fost generat de un model de limbaj, în câteva secunde. Nimeni n-a privit nicio grădină scriindu-l. Nimeni n-a auzit niciun câine. Trei propoziţii fabricate printr-o traiectorie în spaţiul latent — şi totuşi au funcţionat.

Criticul literar Ion Simuţ, citind aceste rânduri, a formulat o observaţie care ascute argumentul: „Mare parte din creaţia literară umană are aceleaşi defecte precum creaţia IA — mediocrităţile generează simulacre de artă, fără a aduce nimic nou. Creatorii mediocri nu ies din serie, lucrează pe principiul combinatoric al clişeelor, procesează formulări preexistente." Observaţia e decisivă: comparaţia pertinentă nu e între IA şi Kafka — ci între IA şi media producţiei literare. Iar acolo, diferenţa e minimă. Textul IA ne place, adaugă Simuţ, tocmai pentru că „răspunde comodităţii noastre de înţelegere, nu ne produce dificultăţi, recunoaştem produsul, nu ne înstrăinează". Asemenea literaturii mediocre — doar cu o viteză şi o bază de date incomparabil mai vaste. Iar la cerere, IA poate genera şi texte avangardiste sau ermetice — dar tot prin imitaţie. Ceea ce confirmă, din perspectiva criticii literare, diagnosticul pe care spaţiul latent îl oferă din perspectivă tehnologică: maşina imită registre, nu generează viziuni.

Filozoful Daniel Dennett ne ajută să înţelegem de ce. Dennett a propus conceptul de postură intenţională (*intentional stance*): strategia prin care interpretăm comportamentul unui sistem atribuindu-i credinţe, dorinţe şi intenţii — indiferent dacă sistemul le posedă sau nu. Adoptăm postura intenţională faţă de oameni („*vrea să mă convingă*"), faţă de animale („*pisica vrea să iasă*"), şi chiar faţă de termostat („*vrea să menţină temperatura*"). Cu un termostat, postura intenţională e o comoditate. Cu un model de limbaj, devine o capcană — pentru că textul generat e atât de elaborat, atât de fluid, încât postura intenţională se activează cu forţa unei evidenţe. Spunem „*IA crede*", „*IA vrea*", „*IA înţelege*" nu din neglijenţă, ci pentru că creierul nostru nu dispune, în faţa unui text coerent, de altă grilă de lectură. Postura intenţională e grila noastră implicită — şi nimeni nu ne-a învăţat s-o dezactivăm în faţa unei maşini care „*vorbeşte*".

Se mai adaugă un alt strat. Unele texte generate sunt ambigue, opace — iar creierul interpretează opacitatea ca profunzime ascunsă. „*Pare profund, deci ascunde ceva.*" Un vers de Celan e opac întrucât *conţine* mai mult sens decât descifrăm la prima lectură. Dar opacitatea unui text generat are altă sursă: e un artefact al procesului de generare, o zonă din spaţiul latent unde mai multe traiectorii converg, producând o formulare indeterminată. Nu e profunzime — e

indeterminare computațională. Diferența e invizibilă dar totală: în literatura autentică, relectura *descoperă* sens; în textul generat, relectura *proiectează* sens. E diferența dintre a descoperi o comoară îngropată și a îngropa tu însuți o comoară pretinzând apoi că ai descoperit-o.

Și mai e ceva — o iluzie pe care o trăim în fiecare conversație cu aceste sisteme. Modelele actuale reține contextul, se adaptează, par să ne „înțeleagă". *„Mă cunoaște, deci mă înțelege."* Oricine a lucrat o seară cu un model de limbaj știe cât de puternică e impresia. Dar modelul nu te cunoaște — te modelează statistic. Diferența e aceeași ca între un prieten care te cunoaște cu adevărat și un algoritm de recomandare care îți sugerează cărți pe baza cumpărăturilor anterioare: algoritmul poate „nimeri" mai des decât prietenul, dar nu te cunoaște deloc. Iar impresia de a fi cunoscut de o mașină e poate cea mai seducătoare dintre iluzii — pentru că răspunde unei nevoi umane profunde de recunoaștere.

De ce toate aceste mecanisme sunt atât de eficiente? Motivul e evolutiv. Timp de sute de mii de ani, orice voce articulată aparținea unei conștiințe. Am dezvoltat o „hipersensibilitate la agenție": detectăm intenție peste tot, chiar și acolo unde nu există — în nori care „par triști", în vânturi care „gem". Acum, pentru prima dată în istoria speciei, există voci articulate fără conștiință. Iar noi nu avem echipamentul cognitiv pentru a le detecta. O față artificială ne deranjează — există o „vale stranie" (*uncanny valley*), o zonă de apropiere în care robotul aproape-uman provoacă repulsie. Dar un text artificial nu provoacă nicio alarmă. Nu există echivalent al „văii stranii" în limbaj. Fluența lingvistică e o suprafață fără fisuri pe care creierul o parcurge fără suspiciune.

Ce putem face? Singura protecție reală nu e intuiția (ne trădează), nici informarea abstractă (dacă vă spun „veți fi păcăliți", tot veți fi) — ci înțelegerea mecanismului. Când știi că fluența e rezultatul optimizării, nu al inteligenței, nu poți elimina iluziile — sunt prea adânc înrădăcinate — dar poți să le recunoști când operează.

Iar acum, amintiți-vă emoția de la început — fragmentul despre grădină. Acum știți ce s-a petrecut: euristica fluenței, circuitele empatice, iluzia profunzimii. Știți toate astea — și totuși, dacă l-ați reciti, emoția ar reveni. De la ELIZA în 1966 la modelele de limbaj în 2026, mecanismul e același — doar puterea iluziei a crescut exponențial. Iluziile cognitive nu se dezactivează prin cunoaștere. Se recunosc. Iar recunoașterea e primul pas spre discernământ.

4. Cât de mult din ceea ce un model de limbaj „scrie" este cu adevărat „nou" — și cât este recombinare? Poate IA să fie cu adevărat „originală"?

- Originalitatea este condiția fără de care literatura nu e literatură. Un text poate fi corect, fluent, chiar elegant, și totuși nul literar dacă nu aduce ceva ce nu exista înainte. Ezra Pound condensa această exigență într-un singur imperativ: *„Make it new!"* — fă-l nou. Nu mai bun, nu mai frumos, nu mai corect. Nou.

Pentru a judeca dacă IA satisface acest imperativ, revenim la spațiul latent. Modelul poate ajunge nu doar în punctele existente (textele reale din corpusul de antrenament), ci și între aceste puncte. Matematic, operația se numește interpolare — producerea unui punct intermediar între doi sau mai mulți vectori existenți. Un text generat poate să nu fie identic cu niciun text din corpusul de antrenament și totuși să fie derivat integral din texte existente, la fel cum portocaliul e derivat din roșu și galben.

Aceasta este sursa confuziei: textul generat pare nou (nu-l găsești nicăieri în bibliotecile lumii), dar nu este nou în sensul lui Pound. Este o combinație inedită de elemente preexistente — nu o creație care vine din afara repertoriului existent.

Distincția este subtilă dar fundamentală, și merită o analogie din lumea pe care cititorii acestei reviste o cunosc cel mai bine.

Imaginați-vă un caleidoscop. Introduceți câteva cioburi colorate. La fiecare rotire, oglinda produce o configurație nouă, pe care n-ați mai văzut-o — simetrică, frumoasă, unică. Puteți roti de un milion de ori fără a obține exact aceeași configurație. Și totuși, fiecare configurație e compusă din aceleași cioburi. Noutatea e combinatorie, nu substanțială. Caleidoscopul produce varietate infinită din materie finită — dar nu produce niciodată un ciob care nu era deja acolo.

Un model de limbaj e un caleidoscop de complexitate vertiginoasă — cu miliarde de „cioburi" (tipare lingvistice) și o „oglindă" (arhitectura neuronală) capabilă de combinații pe care nicio minte umană nu le-ar putea calcula. Varietatea e imensă. Dar substanța rămâne cea furnizată de textele umane.

Se deschide aici o problemă pe care nici estetica, nici dreptul nu au rezolvat-o încă. Fiecare text generat de IA este, în sens matematic riguros, un derivat al tuturor textelor de antrenament — și al niciunuia în particular. Nu e plagiat (nu reproduce un text specific), dar nu e nici original (nu vine din afara repertoriului existent). Spațiul latent funcționează în acest sens asemenea unui palimpsest digital: fiecare text antrenat a lăsat o urmă, dar urmele sunt dizolvate și anonimizate. Nu mai poți reconstitui cine a scris ce, nu mai poți identifica intențiile originale. Rămân doar tiparele. Cine e autorul unui text generat? Utilizatorul, care a formulat promptul? Compania, care a construit modelul? Milioanele de scriitori ale căror texte au alimentat antrenamentul, fără a fi fost consultați? Legile actuale ale dreptului de autor presupun un autor identificabil și o operă originală. Textele generate nu satisfac niciuna dintre aceste condiții.

Să comparăm cu ceea ce face un creator uman. Când Kafka scrie prima frază din *Metamorfoza* — „Într-o bună dimineață, când Gregor Samsa se trezi în patul lui, după o noapte de vise zbuciumate, se pomeni metamorfozat într-o gânganie înspăimântătoare" — el nu interpolează între texte despre insecte și texte despre alienare. Produce o ruptură care nu era conținută în spațiul textelor anterioare. Gregor Samsa nu e un punct intermediar pe nicio hartă — e un continent nou. La fel, când Eminescu scrie *Luceafărul*, nu produce o interpolare între basm și poezie romantică — creează o operă care transcende ambele surse și care generează un spațiu poetic fără precedent. Când Borges inventează *Biblioteca din Babel*, creează un concept care nu exista și care devine instrument de gândire pentru toți cei care vin după.

Aceste acte presupun o ieșire din spațiul latent — producerea a ceva ce nu se află pe hartă, ci o extinde. Iar această ieșire e, prin construcție, imposibilă pentru un model de limbaj. Modelul e prizonier al hărții pe care o traversează. Originalitatea sa este combinatorie (configurații noi ale

acelorași elemente), nu ontologică (elemente noi care nu existau).

Diferența se formulează și altfel: diferența dintre surpriză și revelație. Textul IA poate surprinde — prin asocieri neprevăzute. Dar surpriza e un efect combinatoric. Revelația e altceva: ea aduce la lumină ceva ce nu exista înainte de actul scrierii, ceva ce schimbă modul în care privești lumea *după* ce l-ai citit. Proust descria exact acest efect: *„O adevărată călătorie a descoperirii nu constă în a căuta noi peisaje, ci în a avea ochi noi."* Revelația literară ne dă ochi noi. Caleidoscopul ne dă peisaje noi — dar cu aceiași ochi.

Cineva ar obiecta: și scriitorii umani se bazează pe ce au citit și trăit. Este adevărat. Niciun creator uman nu operează *ex nihilo*. T.S. Eliot o spunea: *„Poeții imaturi imită; poeții maturi fură."* Și adăuga: *„Poetul bun transformă furtul în ceva diferit."* Diferența nu e în prezența materialului anterior, ci în natura procesului. Poetul fură și transformă prin experiență întrupată, conștiință, necesitate, risc. Modelul îl transformă prin calcul probabilistic. Poetul „fură" și transformă — mașina recombină fără a transforma. Una este alchimie; cealaltă este amestecare.

Rilke scria în *Scrisorile către un tânăr poet*: *„Întrebați-vă în ceasul cel mai liniștit al nopții: trebuie să scriu?"* Originalitatea vine din această necesitate — din imposibilitatea de a nu scrie ceea ce scrii, din urgența interioară care face ca tocmai acest text, și nu altul, să fie singurul posibil. Un model de limbaj nu are necesitate. Poate genera orice text cu egală ușurință — ceea ce înseamnă că niciun text nu e, pentru el, necesar. Și fără necesitate, nu există originalitate — există doar varietate.

Du Sautoy formulează acest lucru cu precizie: „*Creativitatea e legată intim de mortalitate, iar mortalitatea e profund înscrisă în definiția ființei umane. [...] Conștiința de a fi muritor e unul din prețurile pe care le plătim pentru conștiință.*" Creativitate, mortalitate, conștiință — cele trei formează un nod inseparabil. Scriitorul scrie din această condiție: timpul lui e limitat, experiența lui e unică și irecuperabilă, fiecare alegere exclude alte alegeri pe care nu le va mai putea face. Mașina nu plătește acest preț. Nu are conștiință, deci nu are conștiința morții, deci nu are urgența de a crea. Produce text fără a risca nimic — nici timp, nici identitate, nici dispariție.

Niciun text nu ilustrează mai bine acest nod decât câteva rânduri ale lui Georges Perec:

„Scriu: scriu pentru că am trăit împreună, pentru că am fost unul dintre ei, o umbră printre umbrele lor, un corp aproape de corpurile lor; scriu pentru că și-au lăsat amprenta de neșters asupra mea, iar urma acesteia este scrisul: memoria lor a murit față de scris; scrisul este memoria morții lor și afirmarea vieții mele."

Perec scria din rană. Fiecare cuvânt al său poartă amprenta unei experiențe trăite — pierderea părinților în Holocaust, copilăria fracturată, corpul printre corpuri. Niciun model de limbaj nu va genera vreodată o astfel de frază, pentru că ea nu vine din limbaj — vine din viață. Iar viața, spre deosebire de limbajul despre viață, nu se află în spațiul latent.

Poate IA să fie cu adevărat originală? Nu — și nu din cauza unei deficiențe tehnice corectabile, ci din cauza naturii procesului. Interpolare mai sofisticată rămâne interpolare. Un caleidoscop mai complex rămâne caleidoscop. Iar originalitatea — cea care justifică existența literaturii — nu este o interpolare mai reușită, ci o ieșire din spațiul a ceea ce exista deja. Această ieșire presupune o ființă care trăiește în lume, nu un sistem care procesează reprezentări ale lumii.

5. Cum poate un scriitor, un critic literar sau un cititor avizat să „vadă" diferența dintre un text uman și un text generat? Există criterii concrete?

- Diferența nu se vede la nivelul propoziției — la acest nivel, textul generat poate fi impecabil. Se vede la niveluri pe care lectura rapidă le-a atrofiat, dar pe care un cititor avizat le poate reactiva. Putem recurge la mai multe criterii verificabile de oricine.

Testul relecturii. Citiți un text de trei ori. Dacă la a doua și a treia lectură descoperiți straturi noi — ambiguități productive, ecouri interne, aluzii care se dezvăluie — textul e probabil uman. Dacă la a doua lectură simțiți sațietate — totul a fost consumat — textul poartă semnătura generării. Un text uman e multistratificat; un text generat e o traiectorie unidimensională prin spațiul latent. Roland Barthes distingea între texte „lizibile" și texte „scriptibile": textul lizibil se consumă, textul scriptibil se recreează la fiecare lectură. Textele IA sunt scriptibile la minimum și lizibile la maximum — se consumă integral la prima lectură.

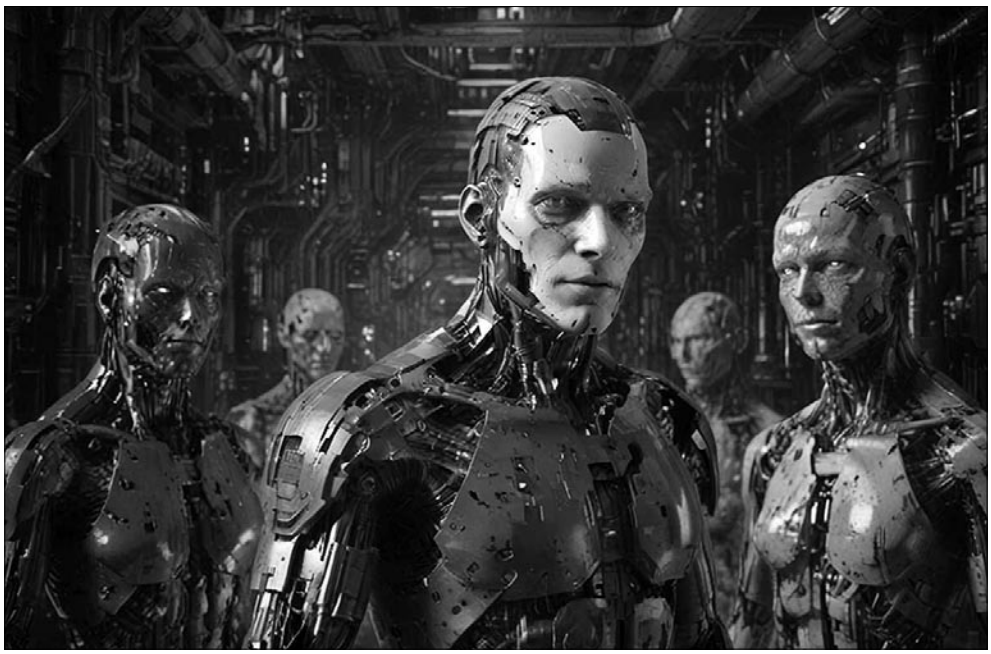
Prezența riscului estetic. Marile opere literare implică un risc — formal, tematic, moral. Riscul este vizibil: simți că scriitorul a mers într-o direcție pe care nu o controla complet, că s-a expus, că putea eșua. Beckett risca reducerea limbajului la bâlbâială. Emily Dickinson risca ermetismul total. Joyce risca ilizibilitatea. Virginia Woolf nota în jurnalul ei: *„Scrisul e ca și cum ai merge pe o scândură deasupra prăpastiei."* Această prăpastie — posibilitatea concretă a eșecului — este semnătura invizibilă a creației autentice. Când o simți în text, știi că cineva a fost acolo. Textele generate de IA nu conțin risc — modelul nu are nimic de pierdut. Absența riscului se simte: textul poate fi impecabil dar inert, corect dar fără tensiune, competent dar fără curaj.

Convergența sau deviația. Textele generate de IA tind structural spre ceea ce te aștepți — spre mediana gustului, spre formularea consensuală, spre ce „sună bine" fără a deranja. Fiind prin construcție produsul automatizării, ele confirmă rutina, nu o rupe. Literatura autentică face exact opusul rutinei: îți strică „pilotul automat". Ea ia lucrurile pe care le credeai cunoscute și le arată dintr-un unghi neașteptat, astfel încât să nu le mai poți trece cu vederea. Te obligă să încetinești, să privești din nou, să simți altfel. În loc să confirme obișnuința, o sparge: ceea ce părea banal devine din nou ciudat, viu și demn de atenție. Când Camus deschide *Străinul* cu „Azi a murit mama. Sau poate ieri, nu știu", acest ton plat în fața morții ne scoate din așteptările obișnuite despre doliu și ne obligă să privim altfel ceea ce credeam evident.

Vocea. Fiecare scriitor autentic are o voce — un mod de a fi în limbaj care este inconfundabil, ireductibil, imposibil de imitat perfect. Vocea nu este stil în sens tehnic (alegerea anumitor figuri de stil sau structuri sintactice) —este prezența unei persoane în text, cu tot ce implică aceasta: temperament, ritm interior, obsesii, limitări. Când citești o pagină de Caragiale, știi că e Caragiale — nu recunoști un procedeu, recunoști o *prezență*. Pentru Flaubert, stilul este o manieră absolută de a vedea lucrurile. Nu de a *scrie* — de a *vedea*. Un model de limbaj poate imita stilul lui Flaubert, dar nu privirea

lui Flaubert. Şi tocmai de aceea versatilitatea totală — capacitatea de a scrie „în stilul" oricui — e semnătura absenţei vocii. Cine poate scrie în toate stilurile nu are niciunul. Vocea presupune o constrângere — scriitorul scrie aşa pentru că nu poate altfel. Versatilitatea totală este semnătura absenţei vocii.

Relaţia cu tăcerea. Într-un text uman, ceea ce nu se spune este la fel de important cu ceea ce se spune. Tăcerile, pauzele, elipsele, spaţiile albe sunt alegerile negative ale autorului — punctele în care a decis să se oprească, să lase loc, să nu spună. Hemingway transforma principiul în metodă:



„Dacă un prozator ştie destul despre ce scrie, poate omite lucruri pe care le cunoaşte." Teoria aisbergului: şapte optimi sub apă. Un model de limbaj nu are relaţie cu tăcerea. Nu omite — generează. Nu alege să nu spună — spune tot ce distribuţia probabilistică îi permite. Un cititor format poate simţi diferenţa: în textul uman, tăcerea apasă; în textul generat, tăcerea e goală.

Dincolo de criteriile de lectură, rămâne o întrebare culturală mai vastă. Dacă nu mai putem distinge uşor între producţii umane şi artificiale, ce se întâmplă cu noţiunea de autenticitate? Într-o piaţă literară inundată de texte generate, cum certificăm provenienţa unei opere? Şi mai profund: dacă generaţia următoare creşte citind texte fără a şti dacă au fost scrise de un om sau fabricate de o maşină, ce se întâmplă cu formarea gustului literar, cu educaţia sensibilităţii, cu transmiterea patrimoniului? Un tânăr cititor care nu distinge între Eminescu şi „în stilul lui Eminescu" nu pierde doar o informaţie — pierde o relaţie cu o conştiinţă umană, o voce situată în istorie. Patrimoniul literar riscă să fie dizolvat în spaţiul latent — nu distrus, ci diluat până la indistincţie.

Suntem condamnaţi la confuzie? Nu. Dar discernământul trebuie cultivat. Cele cinci criterii nu sunt instrumente automate, ci practici de lectură atentă pe care oricine le poate dezvolta. S a spus, pe bună dreptate, că omenirea pare atinsă de o adevărată epidemie în chiar facultatea care o defineşte cel mai mult: folosirea cuvintelor. Diagnosticul, formulat în anii '80, rămâne profetic. Singurul antidot este acelaşi: exactitatea, vizibilitatea, multiplicitatea – calităţile unei lecturi care refuză să se lase păcălită de suprafaţă.

6. Pornind de la cunoaşterea mecanismului intern, cum vedeţi viitorul relaţiei dintre IA şi literatura umană? Ce sfat concret aţi da unui scriitor care se întreabă dacă ar trebui să folosească aceste instrumente?

- Întrebările despre viitor sunt speculative. Dar cunoaşterea mecanismului permite eliminarea a două extreme care domină dezbaterea publică: entuziasmul naiv („IA va democratiza creaţia!") şi panica apocaliptică („IA va distruge literatura!"). Ambele pleacă de la neînţelegerea mecanismului. Între ele se află un teritoriu mai auster — cel al lucidităţii.

Câteva evoluţii sunt certe. Cifrele din 2025 confirmă că nu mai vorbim despre un scenariu ipotetic. Într-un sondaj BookBub pe peste 1.200 de autori, aproape jumătate (45%) declară că folosesc IA cel puţin ocazional în procesul de scriere. Proporţia urcă la 61% într-un raport Gotham Ghostwriters pe aproximativ 1.500 de profesionişti ai scrisului, dintre care un sfert (26%) recurg la IA zilnic — mai ales pentru documentare (81%), marketing, structurare şi editare. Doar 7% publică text generat fără editare substanţială. La ficţiune, adopţia e mai prudentă: 42% (Gotham) sau 33% printre romancierii publicaţi din Regatul Unit, potrivit unui studiu realizat de Universitatea Cambridge pe 258 de autori.

Dar cifrele de adopţie nu spun totul. Acelaşi studiu Cambridge relevă o anxietate profundă: 51% dintre romancierii britanici intervievaţi estimează că IA va ajunge să le înlocuiască integral munca, 85% anticipează scăderi ale veniturilor, iar 39% raportează deja un impact negativ asupra câştigurilor din cauza concurenţei textelor generate. Mai semnificativ: 98% dintre profesioniştii chestionaţi de Gotham exprimă îngrijorări majore legate de drepturile de autor, de halucinaţiile factuale ale modelelor şi de inundarea pieţei editoriale. Peisajul e polarizat: pe de o parte, scriitori care raportează câştiguri de productivitate de peste 30%; pe de alta, scriitori care percep IA drept o ameninţare existenţială — în special în genurile comerciale (roman sentimental, thriller), unde formula narativă e mai uşor de reprodus algoritmic.

Dincolo de statistici, aceste cifre confirmă ceea ce mecanismul pe care l-am descris permite să anticipăm: IA nu ameninţă literatura prin calitate, ci prin volum şi viteză. Şi tocmai de aceea, pentru a naviga în acest peisaj, nu e suficient să ştii dacă să foloseşti IA — trebuie să înţelegi ce face atunci când o foloseşti.

Modelele vor deveni mai performante, textele generate mai greu de distins, volumul de text generat va creşte exponenţial. Piaţa literară va fi inundată de texte fabricate —

fenomenul a și început, cum arată exemplele din preambulul acestei anchete: 200 de romane într-un an, un roman în 45 de minute. Dar inundația de text nu e o înflorire a literaturii. Când textele se înmulțesc fără experiență umană în spate, nu devin mai literare — devin zgomot.

Văd patru scenarii care vor coexista, nu se vor succeda.

Primul: diluția sau inflația. Edituri asaltate, platforme inundate, capacitatea de filtrare depășită. Nu victoria literaturii IA, ci degradarea ecosistemului literar.

Al doilea: segregarea. Circuite separate — producție automată pentru consum de masă, literatură autentică certificată prin edituri, reviste și premii cu garanții de autorship. Ca alimentele artizanale față de cele industriale: pentru publicuri diferite, la prețuri diferite.

Al treilea: hibridizarea. Scriitori folosind IA ca instrument. Risc specific: dacă toți folosesc aceleași modele, textele vor converge stilistic, ca niște pictori care ar folosi toți aceeași paletă limitată.

Al patrulea, cel care mă interesează cel mai mult: reacția umanistă. Exact cum fotografia a eliberat pictura de obligația reprezentării fidele și a deschis calea spre impresionism; exact cum industrializarea a provocat mișcarea *Arts and Crafts* — literatura generativă ar putea provoca o nouă conștiință a ceea ce e ireductibil uman în actul de a scrie.

Paradoxul e acesta: tocmai ceea ce face IA imposibil de egalat în eficiență — viteza, volumul, absența oboselii — este și ceea ce îi lipsește structural. O entitate fără finitudine trăită, fără biografie, fără frica dispariției, poate simula forme creative, dar nu poate atribui valoare producțiilor sale. Valoarea pe care noi o acordăm operelor vine din recunoașterea condiției noastre comune de ființe muritoare — și această dimensiune existențială rămâne, deocamdată și probabil pentru totdeauna, specific umană.

Walter Benjamin se întreba în 1936 ce se întâmplă cu opera de artă în epoca reproducerii mecanice — și constata pierderea „aurei". Un secol mai târziu, întrebarea se pune la alt nivel: nu reproducerea, ci producerea mecanică a operei. Și totuși, lecția rămâne: ceea ce se pierde prin mecanizare devine, prin contrast, cel mai prețios. Aura nu dispare — devine rară.

Ce sfat concret aș da unui scriitor?

Nu-mi stă în fire să dau sfaturi unui scriitor — creația e un teritoriu în care fiecare își cunoaște drumul mai bine decât orice observator extern. Pot însă, pornind de la ceea ce am explicat despre mecanism, să atrag atenția asupra unor aspecte pe care le consider esențiale.

Primul privește înțelegerea instrumentului. Un scriitor care folosește IA fără a ști ce se petrece sub capotă se expune, fără să-și dea seama, unui risc specific: acela de a confunda ce generează mașina cu propria gândire. Cioran scria: „*Orice lucru pe care mi-l însușesc mă sărăcește.*" Însușirea textului produs de o mașină e o formă de sărăcire creativă — nu materială, ci interioară. Scriitorul, poetul, creatorul au totul de câștigat înțelegând conceptual (nu neapărat tehnic) cum funcționează sistemul pe care îl folosesc.

Al doilea privește granița dintre instrument și substituție. IA este un instrument excelent de documentare, de verificare, de testare a coerenței unui argument. Dar în momentul în care un creator îi deleagă alegerea cuvântului just, construcția vocii, decizia de a spune sau de a nu spune — în acel moment creatorul nu mai este prezent în text. Iar un text în care autorul nu se află este, indiferent de calitate, un text fără autor — adică exact ceea ce mașina produce și singură. Rilke formula cu precizie această condiție: a scrie doar ceea ce trebuie scris, ceea ce nu poți să nu scrii. Dacă un scriitor poate delega scrierea unui algoritm, probabil acel text nu trebuia scris de el.

Al treilea e poate cel mai important și privește ceea ce mașina nu poate face. Experiența trăită, curajul estetic, vocea unică, relația cu tăcerea, interogarea realității — acestea nu sunt calități amenințate de IA. Sunt, dimpotrivă, calitățile pe care prezența IA le face mai vizibile și mai prețioase. Un scriitor nu trebuie să se apere de IA — trebuie să cultive în el ceea ce IA îi revelează, prin contrast, drept ireductibil.

De-a lungul vieții mele profesionale, am trăit mai multe revoluții tehnologice — de la cartelele perforate la calculatorul personal, de la Internet la inteligența artificială. De fiecare dată, noua tehnologie a provocat fascinație și teamă. De fiecare dată, cei care au înțeles mecanismul au putut transforma instrumentul în aliat, iar cei care s-au lăsat fascinați de aparențe au fost deposedați de propriile competențe.

7. Aveți în pregătire o carte dedicată funcționării interne a IA generative. Ce ar descoperi un scriitor sau un poet citind-o? Ce schimbă înțelegerea mecanismului în raport cu simpla utilizare?

- De doi ani lucrez la o carte intitulată *Au cœur de l'IA générative. Fonctionnement, espace latent et illusions cognitives*, care va apărea la FYP Éditions în iunie 2026. Cele trei cuvinte-cheie ale subtitlului — funcționare, spațiu latent, iluzii cognitive — sunt cele trei fire pe care le-am urmat în aceste răspunsuri.

De ce o carte întreagă? Și de ce ar citi-o un scriitor?

Există o diferență fundamentală între a conduce o mașină și a înțelege ce se petrece sub capotă. Șoferul experimentat conduce bine fără a ști cum funcționează motorul. Dar când motorul face ceva neașteptat, șoferul care nu înțelege mecanismul e neajutorat: nu poate distinge între simptom și cauză, între benign și periculos.

Cu IA generativă suntem în această situație. Milioane de oameni „conduc" zilnic aceste sisteme. Puțini înțeleg ce se petrece sub capotă. Iar când sistemul produce un text care pare profund sau unul care pare absurd, utilizatorul care nu înțelege mecanismul nu poate distinge între aparență și realitate. E la mila impresiei — și impresia e guvernată de iluziile cognitive pe care le-am descris.

Am ales ca fir unificator al celor douăsprezece capitole conceptul de spațiu latent — acest univers matematic invizibil — pentru că el este cheia care deschide toate ușile: de la modelele de limbaj la generarea de imagini, de la creația muzicală la producția video. Același concept, același principiu, aceleași iluzii.

Ce ar descoperi un scriitor citind-o?

Ar descoperi o demistificare fără trivializare. Cartea nu spune „IA e doar statistică" — ar fi o simplificare. Nici „IA e o nouă inteligență" — ar fi o mistificare. Spune: iată ce se petrece efectiv, pas cu pas, strat cu strat. Și lasă cititorul să judece — cu instrumente, nu cu impresii.

Goethe reformula o veche maximă a lui Epictet: oamenii nu sunt deranjați de lucruri, ci de ideile pe care și le fac despre lucruri. Cu IA se întâmplă exact asta. Cartea își propune să înlocuiască ideea (de regulă greșită) cu înțelegerea — care e

întotdeauna mai nuanțată și mai puțin spectaculoasă decât iluzia.

Apoi, ar descoperi o explicație a propriei vulnerabilități cognitive. Un scriitor care parcurge capitolele despre iluzii nu va înceta să fie impresionat de textele IA, dar va putea recunoaște momentul în care e impresionat și va putea pune între paranteze impresia. E diferența dintre a fi orbit de un reflector și a ști că e un reflector — ești tot orbit, dar nu mai crezi că vezi soarele.

Și poate cel mai important: ar descoperi o revalorizare a propriei meserii. Prezența IA în câmpul cultural poate, paradoxal, să intensifice simțirea a ceea ce rămâne ireductibil uman în actul creației — la fel cum, în pictură, apariția fotografiei nu a ucis tabloul, ci l-a eliberat, revelându-i esența. Dar pentru a se produce această revalorizare, trebuie mai întâi să înțelegem ce e IA — nu ce pare, nu ce promite, nu ce ne temem să fie, ci ce este. Cartea pe care am scris-o nu face pe cititor mai competent în utilizarea IA. Îl face mai lucid — și, prin această luciditate, mai conștient de valoarea a ceea ce el însuși produce când scrie, când citește, când gândește. Nu ceea ce știm ne face mai buni sau mai slabi — ci ceea ce facem cu luciditatea dobândită.

René Char scria: „*Luciditatea este rana cea mai apropiată de soare.*" Într-o epocă de simulacre, luciditatea rămâne cea mai înaltă formă de rezistență. A ști ce e mașina, a ști ce poate și ce nu poate, a ști de ce ne păcălește — și a continua, în ciuda acestei cunoașteri, să scrii cu mâna ta, din viața ta, cu vocea ta. Nu pentru că mașina scrie mai prost. Ci pentru că tu, scriind, ești acolo. Iar mașina, oricât de performantă, nu e nicăieri.

* Ioan Roxin a absolvit ca șef de promoție Facultatea de Cibernetică, Statistică și Informatică Economică din București (1980). Un an după susținerea tezei de doctorat, o bursă a Guvernului francez îl aduce în 1991 la Institutul Național de Științe Aplicate (INSA) din Lyon, pentru studii aprofundate în ingineria informaticii. Aici descoperă neuronul artificial și rețelele de neuroni — pe atunci de tip Hopfield (John Hopfield va primi, alături de Geoffrey Hinton, premiul Nobel pentru fizică în 2024, tocmai pentru contribuțiile fundamentale la bazele învățării automate). Formarea în cibernetică îi permite să asimileze cu ușurință aceste concepte noi — drumul de la sistemele cu conexiune inversă la rețelele neuronale era, pentru un cibernetician, mai scurt decât pentru alții. Rămâne în Franța, unde parcurge treptele universitare — conferențiar, apoi profesor — la Universitatea Franche-Comté, astăzi Universitatea Marie et Louis Pasteur. Din septembrie 2024 este profesor emerit la aceeași universitate.

Între 2014 și 2023 a condus laboratorul de cercetare interdisciplinar ELLIADD (Édition, Langages, Littératures, Informatique, Arts, Didactique, Discours), reunind peste 160 de cercetători. Profesor invitat în Brazilia, Canada, Finlanda, Japonia, Liban, Maroc, Mexic, Tunisia și Vietnam, i-a fost decernat ordinul Palmes Académiques — în grad de Cavaler, apoi de Ofițer — prin decretul Prim-ministrului Republicii Franceze.

Inițiator al mai multor specializări universitare în multimedia, semnează volumul de referință *Multimédia. Les fondamentaux — Introduction à la représentation numérique* (împreună cu Daniel Mercier, 2004). Cartea sa *Au cœur de l'IA générative. Fonctionnement, espace latent et illusions cognitives*, programată la FYP Éditions în iunie 2026 — cu o ediție în limba română în pregătire —, stă la baza răspunsurilor din acest interviu.

Referințe

Camus, A. (1971). *L'Étranger*. Paris : Gallimard, coll. Folio.

Char, R. (2007). *Feuillets d'Hypnos*. Paris : Gallimard, coll. Espoir.

Chklovski, V. (2008). *L'Art comme procédé*. Paris : Allia.

Cioran, E. M. (1977). *Précis de décomposition*. Paris : Gallimard.

Collett, C. (2025). *The Impact of Generative AI on the Novel*. Minderoo Centre for Technology and Democracy, University of Cambridge. https://doi.org/10.17863/CAM.122470.

Eliot, T. S. (1997). *The Sacred Wood. Essays on Poetry and Criticism*, Faber & Faber

Flaubert, G. (1980). *Correspondance Flaubert*. Paris : Gallimard, coll. Folio.

Gotham Ghostwriters and WOBS LLC. *A.I. and Writing Profession. A Comprhensive Survey & Analysis*. https://gothamghostwriters.com/wp-content/uploads/2025/11/AI-Writer-Survey.pdf

Hemingway, E. (1972). *Mort dans l'après midi*. Paris : Gallimard, coll. Folio.

Kafka, F. (2023). *The Metamorphosis*, Happy Hour Books.

Montaigne, M. de. (2019). *Essais*. Paris : Bouquins.

Pound, E. (1934). *Make it new*. Faber and faber limited, London, https://archive.org/details/in.ernet.dli.2015.185999/mode/2up

Proust, M. (2024). *À la recherche du temps perdu*. Paris: Gallimard.

Rilke, R. M. (2002). *Lettres à un jeune poète*. Grasset.

Robertson, C. (2025). How Authors Are Thinking About AI (Survey of 1,200+ Authors). BookBub. https://insights.bookbub.com/how-authors-are-thinking-about-ai-survey/.

Sautoy, M. du. (2022). *Le code de la créativité. Comment l'IA apprend à écrire, peindre et penser*. Flammarion.

Woolf, V. (1999). *Journal d'un écrivain*. Éditions 10/18.

Interviul și prezentarea au fost realizate de către

VIRGIL DIACONU

Februarie 2026

LEXIC PENTRU CITITORUL LITERAR
Conceptele-cheie ale inteligenței artificiale generative, explicate fără ecuații

Antrenament (*training*) — Procesul prin care un model de limbaj „învață" din texte. Nu e lectură, ci extragere statistică: miliarde de texte sunt parcurse, iar tiparele lor (ce

cuvinte urmează după alte cuvinte, ce structuri sunt frecvente, ce asocieri revin) sunt înregistrate sub formă de parametri numerici. Modelul nu reține textele — reține regularitățile lor.

Corpus de antrenament — Totalitatea textelor pe care un model le-a parcurs în faza de antrenament. Pentru modelele actuale: sute de miliarde de cuvinte — literatură, presă, enciclopedii, conversații, articole științifice, poezie, forumuri. Corpusul determină orizontul modelului: ceea ce nu se află în corpus nu există în spațiul latent.

Distribuție de probabilitate — Pentru fiecare cuvânt pe care modelul urmează să-l genereze, sistemul calculează probabilitatea tuturor cuvintelor posibile. „Geam" poate avea 35%, „fereastră" 20%, „suflet" 3%. Generarea constă în eșantionarea din această distribuție — nu în alegere deliberată, ci în selecție probabilistică.

Efectul ELIZA — Tendința utilizatorilor de a atribui înțelegere și empatie unui program informatic, observată pentru prima dată în 1966 de Joseph Weizenbaum. Programul ELIZA simula un psihoterapeut prin reformulări mecanice, dar utilizatorii se atașau de el și îi atribuiau sensibilitate. Efectul e amplificat exponențial de modelele actuale.

Eșantionare (*sampling*) — Operația prin care modelul „alege" următorul cuvânt dintr-o distribuție de probabilitate. Nu e o decizie — e un tiraj ponderat, similar cu o extragere la loto în care anumite bile sunt mai mari decât altele.

Euristica fluenței (*fluency heuristic*) — Mecanism cognitiv pre-conștient prin care creierul evaluează automat un mesaj ușor de procesat drept mai adevărat și mai valoros. Modelele de limbaj, optimizate pentru fluență, exploatează sistematic această euristică.

Halucinator (*hallucination*) — Termen consacrat (deși imprecis) pentru situațiile în care un model generează informații false cu deplină fluență: referințe bibliografice inexistente, citate inventate, fapte fabricate. Modelul nu „minte" — generează tokenul cel mai probabil indiferent de adecvarea factuală.

Hipersensibilitate la agenție (*hypersensitive agency detection*) — Tendința evolutivă a creierului uman de a detecta intenție și conștiință peste tot, inclusiv acolo unde nu există. Adaptativă în mediul natural (mai bine să detectezi un prădător inexistent decât să ratezi unul real), devine sursă de iluzie în fața textelor generate de IA.

Interpolare — Operația matematică de producere a unui punct intermediar între două sau mai multe puncte existente. În spațiul latent, generarea unui text „între" Bacovia și Baudelaire e o interpolare: rezultatul pare nou, dar e derivat integral din punctele existente. Opusul creației autentice, care *extinde* spațiul în loc să-l traverseze.

Model de limbaj (*Large Language Model — LLM*) — Sistem informatic antrenat pe cantități masive de text, capabil să genereze text nou prin predicție probabilistică. Exemple: Claude (Anthropic), DeepSeek (DeepSeek AI), Gemini (Google), GPT-5 (OpenAI), Grok (xAI de Elon Musk), Mistral (Mistral AI). „Mare" (*large*) se referă la numărul de parametri: de ordinul miliardelor.

Parametri — Valorile numerice (ponderi matematice) ajustate în cursul antrenamentului. Ele codifică tiparele statistice ale limbajului. Un model actual poate avea sute de miliarde de parametri. Sunt echivalentul computațional al „memoriei" modelului — dar o memorie fără amintiri, fără afect, fără experiență.

Postură intenționlă (*intentional stance*) — Concept introdus de filozoful Daniel Dennett: strategia prin care interpretăm comportamentul unui sistem atribuindu-i credințe, dorințe și intenții, indiferent dacă le posedă. Adoptăm postura intențională și față de animale („pisica vrea să iasă"), și față de termostate. Față de un model de limbaj, postura devine o sursă sistematică de iluzie.

Prompt — Textul furnizat de utilizator unui model de limbaj, pe baza căruia modelul generează răspunsul. Promptul funcționează drept punct de plecare al navigării în spațiul latent. Calitatea promptului influențează direct rezultatul generat.

Rețea neuronală (*neural network*) — Arhitectura informatică inspirată (foarte distant) de structura creierului biologic. Straturile succesive de „neuroni artificiali" (funcții matematice) transformă datele de intrare pas cu pas. Modelele actuale de limbaj folosesc o arhitectură specifică numită *Transformer* (2017).

RLHF (*Reinforcement Learning from Human Feedback*) — Etapă a antrenamentului în care răspunsurile modelului sunt evaluate de oameni, iar modelul e ajustat pentru a produce răspunsuri preferate. Acest proces optimizează fluența, claritatea și conformitatea — și contribuie direct la euristica fluenței descrisă mai sus.

Spațiu latent (*latent space*) — Spațiul matematic de dimensionalitate înaltă în care sunt reprezentate intern toate datele din antrenament. Fiecare cuvânt, expresie sau fragment de text devine un vector (un punct cu coordonate) în acest spațiu. Textele similare sunt aproape, cele diferite sunt departe. Generarea de text = navigare prin acest spațiu. „Latent" vine din latină (*latens* — ascuns): spațiul este invizibil pentru utilizator.

Token — Unitatea de bază pe care modelul o procesează. Nu e întotdeauna un cuvânt întreg: poate fi o silabă, un prefix, un semn de punctuație. Generarea se face token cu token — fiecare depinzând de toate cele precedente.

Transformer — Arhitectura de rețea neuronală (propusă în 2017 de cercetători de la Google) pe care se bazează toate modelele actuale de limbaj. Inovația principală: mecanismul de *atenție* (*attention*), prin care modelul poate lua în considerare simultan toate cuvintele anterioare, nu doar cele imediat precedente.

Uncanny valley (*valea stranietății*) — Fenomen observat în robotică: un robot aproape-uman provoacă o reacție de repulsie — e prea asemănător pentru a fi ignorat, prea diferit pentru a fi acceptat. În limbaj, acest fenomen nu există: un text generat, oricât de artificial în origine, nu provoacă nicio alarmă cognitivă.

Vector — Reprezentare numerică a unui cuvânt sau text sub forma unui șir de numere (coordonate într-un spațiu matematic). „Tristețe" și „melancolie" au vectori apropiați; „tristețe" și „șurub" au vectori îndepărtați. Vectorii permit modelului să „calculeze" relații semantice — fără a înțelege sensul.