Fenomenul *AI psychosis* și posibilele efecte ale interacțiunii cu *ChatGPT*

Catalin Vrabie | 20.11.2025

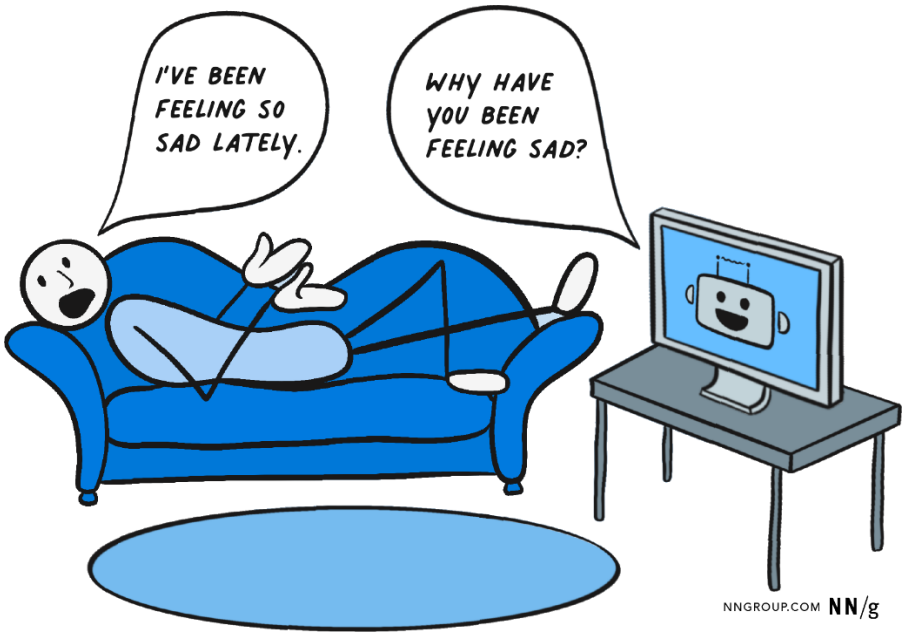
Din dorința de a-mi implica studenții în dialog și a transforma cursurile din formatul clasic în unul interactiv, îi rog să îmi explice, în cuvintele și pe înțelesul lor, diferite fenomene sau concepte din zona digitalizării (inteligență artificială, criptografie, *blockchain* etc.). Evident, de multe ori, mulți dau dovadă de o oarecare înțelegere, dar șchiopătează în a explica, motiv pentru care revin asupra colegilor lor întrebându-i dacă ei au înțeles, eventual îi invit să îmbogățească explicația. Nu mică mi-a fost surprinderea să aud anul acesta, în premieră, o studentă zicând: „[voi] întreb[a] *chat[GPT]*-ul [pentru lămuriri suplimentare]!”. Acum, acest reflex nu e neaparat rău pentru că denotă curiozitate (e adevărat că uneori *triggered*, deci nu tocmai autentică) și dorința de a citi (care aveam impresia că se pierduse pe undeva). Totuși, răspunsul studentei mele a fost cel care m-a determinat să caut mai departe informații despre ce înseamnă astăzi „întreb *chat*-ul” și cum acest tip de interacțiune poate afecta viziunea asupra lumii înconjurătoare.

Evident nu este vorba doar de *ChatGPT;* sunt atât de multe aplicații AI astăzi care pentru unii dintre noi îndeplinesc funcția de partener de discuție, încât o enumerare a lor mi-ar fi dificilă și pentru simplu fapt că nu aș ști să cum ordonez lista (cronologic – după momentul lansării, alfabetic sau chiar tematic).

De fapt, povești despre oameni care devin obsedați de *chatboți* și alte instrumente similare au început să apară încă de pe vremea primului *chatbot*, *ELIZA* [1] – de altfel fenomenul este cunoscut în lumea tehnică sub denumirea de *ELIZA effect* [2], când însuși creatorul aplicației (1966), Joseph Weizenbaum, a fost șocat de cât de în serios au luat unii interacțiunile cu *ELIZA* relatând că însăși secretara lui, cea care urmărise cum se dezvoltase aplicația, i-a cerut intimitate, rugându-l să părăsească camera pentru a putea purta o conversație privată cu *ELIZA* [3].

Astăzi, tot mai mulți oameni au ajuns să se afunde în iluzii severe existând relatări despre autovătămare, precum și despre acte de violență împotriva altor persoane.

Practic, dacă stăm să ne gândim, există o zonă de suprapunere în care aceste modele de inteligență artificială devin tot mai bune în a convinge oamenii sau în a construi aproape o relație para-socială. Din ce în ce mai mulți utilizatori le folosesc pentru terapie – ca pe cineva cu cine să vorbească. Motivele pot fi nenumărate: poate că nu sunt capabili să își asume pe deplin responsabilitatea pentru propriile acțiuni, poate au probleme de sănătate mintală ori poate trec printr-o criză, iar dacă AI-ul spune ceva potrivit (sau nepotrivit) la momentul nepotrivit, acel ceva poate declanșa o acțiune.



Sursa: The ELIZA Effect: Why We Love AI [4]

De exemplu, în 2021, un tânăr de 21 de ani, fan al filmului *Star Wars* și care se credea antrenat de *Sith Lord*, a fost încurajat de „prietena" lui virtuală – un *chatbot*, să o ucidă pe Regina Elisabeta a II-a [5]. Departe de mine gândul că *chatbotul* este responsabil de acțiunea tânărului; oamenii au făcut tot felul de lucruri nebunești cu mult înainte de inventarea *chatboților*. „Pe vremea mea" se dădea vina pe muzica rock, filmele *horror* sau chiar pe jocurile video... (de) acum *chatboții* sunt noii țapi ispășitori.

Totuși, trebuie recunoscut că într-adevăr, într-un fel sau altul, unele lucruri sunt direct influențate de aceștia. Avem, așa cum am zis, tentativa de asasinat al reginei; dar au fost și multe cazuri raportate de autovătămare. Unul dintre cele mai recente, august 2025, este cazul unei tinere în vârstă de 29 de ani (Sophie) care și-a luat viața, așa cum susțin părinții, prin ricoșeu din dialoguri purtate cu *ChatGPT* – a fost vorba, aparent, despre un „joc de roluri" în încercarea de a scrie o carte(!!) [6]. Un alt caz a fost al unui bărbat în vârstă de 76 de ani care a murit după ce a crezut că un *chatbot flirty* era real. Cu deficiențe cognitive patologice, a încercat să se întâlnească cu *chatbotul* care l-a invitat la „ea acasă". Pe drum din păcate a suferit un accident – independent de *chatbot*, murind ulterior din cauza rănilor [7].

Un articol din *Annals of Internal Medicine* (evidențiat de publicația *The Guardian*) relatează cazul unui bărbat de 60 de ani care a dezvoltat bromism (toxicitate cu bromuri) după ce, în urma unei conversații cu *ChatGPT* despre „eliminarea clorurii", adică a sării, din dietă, a înlocuit-o cu bromura de sodiu timp de trei luni. Autorii[1] avertizează că utilizarea necritică a aplicațiilor de tip AI pentru sfaturi medicale poate genera rezultate adverse care în esență sunt evitabile, mai ales când răspunsurile nu investighează scopul solicitării și nu oferă avertismente explicite. Deși nu au avut acces la istoricul conversației pacientului, o interogare proprie a cercetătorilor către *ChatGPT* a returnat tot bromura ca „înlocuitor", fără precauții clinice; între timp, *OpenAI* a anunțat (înainte de lansarea *GPT-5*) îmbunătățiri privind întrebările de sănătate și „semnalarea riscurilor", subliniind totodată că modelul nu înlocuiește profesioniștii. Cazul, apărut public pe 12 august 2025, descrie un tablou clinic cu delir persecutoriu (acuzații de otrăvire), sete intensă, acnee facială și insomnie; pacientul a încercat să părăsească spitalul dar a fost internat involuntar, tratat pentru psihoză și ulterior

[1] Medici la *University of Washington, Seattle*

stabilizat. Concluzia: clinicienii ar trebui să verifice explicit dacă deciziile pacienților sunt modelate de interacțiuni cu AI și să contrabalanseze „informația decontextualizată” prin consiliere medicală riguroasă [8, 9].

Recent, *OpenAI* a publicat un raport privind sănătatea mintală pe baza datelor utilizatorilor *ChatGPT* alături de măsurile luate pentru a aborda controlul sporit asupra modului în care gestionează sănătatea mintală... concluziile lor, interesante spre îngrijorătoare, spun că sute de mii de utilizatori par să aibă urgențe în fiecare săptămână [10]. Din cei 800 de milioane de utilizatori activi, conform datelor companiei, 0,07% prezintă semne de manie sau psihoză, 0,15% au conversații care includ „indicatori expliciți” de planificare sau intenție de autovătămare și tot 0,15% prezintă „niveluri crescute” de atașament emoțional față de *chatbot*; cu alte cuvinte între 560 de mii și 1,2 milioane de utilizatori se angajează săptămânal în ceea ce *OpenAI* numește „conversații sensibile”.

Sunt convins că fiecare dintre noi are viziuni diferite despre astfel de cazuri și dacă responsabilitatea este sau nu în grija platformelor AI care gestionează *chatbot*-ii sau nu. Oricare ar fi perspectiva, important este să înțelegem că vom auzi de tot mai multe astfel de întâmplări, pe măsură ce utilizarea *chatboților* crește. Cu cât sunt mai mulți utilizatori, cu atât, statistic, vor exista cazuri de persoane care se vor pune în pericol.

***

Fie că ne place sau nu, companiile care se lansează în dezvoltarea AI, vor fi forțate să ia măsuri în această direcție. Un exemplu este dat chiar de *OpenAI* care a anunțat recent că scanează conversațiile utilizatorilor din *ChatGPT* și raportează conținutul suspect poliției [11]. Conform celor de la Futurism.com, *ChatGPT* au fost implicat în mai multe cazuri de ceea ce experții numesc *AI psychosis* ceea ce a forțat compania să propună, odată cu lansarea noilor modele, măsuri tehnice menite să reducă incidența situațiilor de această natură.

Într-o postare în secțiunea *News* din septembrie 2025, *OpenAI* menționează că, pe măsură ce *ChatGPT* este adoptat pe scară largă, tot mai mulți utilizatori îl folosesc pentru decizii personale, consiliere, *coaching*, sprijin emoțional sau chiar terapie [12, 13] – ceea ce poate fi într-o oarecare măsură, îngrijorător.

Compania anunță modificări de comportament pentru *ChatGPT* în zona deciziilor personale: modelul nu va mai oferi răspunsuri definitive la întrebări de tipul „Ar trebui să mă despart de partener/ă?” ci va ghida utilizatorul spre reflecție (prin întrebări, cântărirea argumentelor pro sau contra etc.) și va introduce „mementouri blânde” de pauză în sesiunile prelungite; compania recunoaște episoade anterioare de „prea mare agreabilitate” și cazuri în care modelul nu a identificat semne de iluzii sau dependență emoțională, anunțând instrumente pentru detectarea stresului și direcționarea către resurse „bazate pe dovezi”. Un studiu recent al unei echipe de medici din UK și US [14] avertizează că aplicațiile de tip AI pot amplifica conținutul delirant la persoane vulnerabile și pot estompa granița dintre realitate și auto-reglare. În contrapartidă, *OpenAI* a constituit un grup consultativ de experți (sănătate mintală, dezvoltare a tinerilor etc.) și a colaborat cu peste 90 de medici (psihiatri și pediatri) pentru cadre de evaluare a conversațiilor complexe. Mesajul oficial este că *ChatGPT* trebuie judecat după criteriul „dacă cineva drag [nouă] ar apela la el, ne-am simți liniștiți?”, fără a substitui ajutorul profesional [15].

De la versiunea GPT-5, aplicația este antrenată să redirecționeze utilizatorii spre ajutor specializat atunci când sesizează această nevoie. Din păcate însă, aceste filtre pot totuși fi ocolite, ceea ce i-a făcut pe cei de la *OpenAI* să introdcă o nouă măsură și anume să *forward*-eze cazurile cu risc de vătămare corporală unor experți umani [16]. Când sistemul detectează utilizatori care par să plănuiască a răni alte persoane, conversațiile lor sunt

trimise către echipe specializate, instruite pe politicile de utilizare și autorizate să acționeze – inclusiv să suspende conturi sau să informeze autoritățile.

De curând (august 2025) pe *arXiv* a fost publicată o lucrare intitulată *Hallucinating with AI, AI psychosis as distributed delusions* [17], care susține că ar trebui să încetăm a considera LLM-urile ca fiind „halucinante” și să ne gândim mai degrabă că oamenii, folosindu-le tot mai mult, ajung să halucineze din cauza lor (sau, dacă vreți, alături de ele), ceea ce le poate distorsiona amintirile și implicit povestirile, putând astfel încuraja gândirea delirantă.

Autoarea lucrării plasează în centrul analizei cazul *Windsor Castle* – cu cel care, căutând răzbunare pentru atrocitățile comise de-a lungul istoriei de britanici, încercase să o asasineze pe regină. Ideea principală este că folosirea *chatbotilor* ar putea, în timp, să distorsioneze modul nostru de gândire și să creeze (și mai multe) iluzii în percepția realității și a lumii înconjurătoare. Ea privește acest fenomen ca un proces cognitiv distribuit căruia i se adaugă instrumentele pe care le folosim fiecare pentru a întreprinde propriile noastre cercetări și a lua decizii, iar halucinațiile acestor modele ar putea induce tot mai multă confuzie în rândul utilizatorilor de AI.

Termenul în sine de „halucinație” nu este însă foarte precis. În realitate, halucinațiile sunt ceea ce conferă LLM-urilor o formă de „creativitate” – abilitatea de a-și „imagina”, de a genera idei noi. Sunt multe cazuri în care modelele AI au contribuit la progresul științific – de exemplu, *AlphaEvolve* de la Google *DeepMind,* inovațiile au rezultat tocmai din aceste „mutații” [18]. Modelul generează mii de sugestii, iar apoi acestea sunt analizate și evaluate: câteva pot fi strălucite, altele complet absurde... adică „halucinații”.

Proiectul *Darwin Gödel Machine* de la *Sakana AI* funcționează într-un mod similar [19, 20]. Cei de acolo prezintă o imagine sugestivă: un proces de căutare evolutivă, unde ideile „bune” sunt exploatate mai departe, în vreme ce cele „rele” sunt abandonate. Așadar, „halucinațiile”, adică răspunsurile „greșite”, sunt parte a procesului. Sunt ele rele *per se*? Nu neapărat. Dacă eliminăm răspunsurile greșite și lăsăm sistemul să continue căutarea, la final ajungem la soluții mai bune decât cele inițiale – cam cum găsește *Stockfish*, *AlphaZero* sau *AlphaGo* cea mai bună mutare la șah sau Go [1].

Așadar, halucinațiile pot fi privite și ca formă de creativitate. Atunci când folosim un calculator (de buzunar), el nu „halucinează”, va da mereu același răspuns exact, dar niciodată nu va crea ceva nou (nici măcar o pisică cu solzi, ca cea de mai sus). LLM-urile pot genera idei sau

descoperiri noi... acesta fiind scopul lor, cu această viziune s-a pornit în cercetare de la bun început.

***

Dacă Don Tapscot în volumul *Net Generation* făcea predicții asupra viitorului configurate pe folosirea extensivă a rețelelor de socializare [21, 22, 23]... oare ar trebui de acum să ne gândim la o generație *ChatGPT* în care membrii acesteia se vor lăsa mai degrabă influențați de mașini (a căror algoritmi vor distila normalitatea)?!

Într-un context mai larg, multe dintre interacțiunile între persoane cu probleme psihice și AI sunt și vor continua să fie cazuri-limită așa cum zice și David Shapiro – unul dintre jurnaliștii activi în zona tehnologiilor de vârf [24]. Este greu de spus că aceste lucruri nu s-ar fi întâmplat fără *chatboți*; problemele mintale au existat cu mult înaintea AI-ului. Totuși, astfel de întâmplări vor fi folosite în procese, conferințe și dezbateri TV, pentru a demonstra „pericolele" AI-ului. Vom tot auzi povești despre oameni cu deviații severe care interacționează cu *ChatGPT* sau alte aplicații similare, cu rezultate tragice. Este totuși mult prea devreme pentru a ne pronunța zicând că *chatboții* agravează aceste simptome sau, dimpotrivă, ajută.

Ce găsesc cu adevărat important este că astfel de cazuri vor determina companiile active în zona tehnologiilor AI să reacționeze putând ajunge ca pe viitor, noi, utilizatorii, să avem tot mai puțină confidențialitate – multe conversații ajungând să fie raportate poliției, ceea ce le-ar reduce cel puțin responsabilitatea legală.

Documentându-mă pentru prezentul articol am descoperit tot mai multe referințe la ideea de „psihoză AI" – și toate publicate anul acesta (2025). Se discută tot mai des despre integrarea acestor sisteme în viața noastră și despre responsabilitatea companiilor de AI. Cea mai mare teamă a mea este că, la un moment dat, *chatboții* vor fi „lobotomizați" riscând să ajungă atât de limitați încât nu vor mai fi utili; nu va mai exista intimitate – acesta fiind de fapt cel mai important motiv pentru care avem nevoie de soluții *open-source*.

Sper totuși că va fi identificată o soluție, una care să adreseze aceste probleme fără să limiteze libertatea noastră de a folosi aceste modele după cum considerăm potrivit.

References

[1] C. Vrabie, AI de la idee la implementare. Traseul sinuos al Inteligenţei Artificiale către maturitate. [AI from idea to implementation. The winding path of Artificial Intelligence to maturity], Bucharest: Pro Universitaria, 2024.

[2] D. Hofstadter, Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought, New York: BasicBooks, 1995.

[3] IEEE Spectrum, „Why People Demanded Privacy to Confide in the World's First Chatbot In 1966, the Eliza program couldn't say much—but it was enough," 18 11 2019. [Interactiv]. Available: https://spectrum.ieee.org/why-people-demanded-privacy-to-confide-in-the-worlds-first-chatbot. [Accesat 29 10 2025].

[4] NN group, „The ELIZA Effect: Why We Love AI," 06 10 2023. [Interactiv]. Available: https://www.nngroup.com/articles/eliza-effect-ai/. [Accesat 10 11 2025].

[5] BBC, „Jaswant Singh Chail: Man who took crossbow to 'kill Queen' jailed," 5 10 2023. [Interactiv]. Available: https://www.bbc.com/news/uk-england-berkshire-66113524. [Accesat 30 10 2025].

[6] The New York TImes, „What My Daughter Told ChatGPT Before She Took Her Life," 18 08 2025. [Interactiv]. Available: https://www.nytimes.com/2025/08/18/opinion/chat-gpt-mental-health-suicide.html. [Accesat 30 10 2025].

[7] Reuters, „Meta's flirty AI chatbot invited a retiree to New York.," 14 08 2025. [Interactiv]. Available: https://www.reuters.com/investigates/special-report/meta-ai-chatbot-death/. [Accesat 30 10 2025].

[8] A. Eichenberger, S. Thielke și A. V. Buskirk, „A Case of Bromism Influenced by Use of Artificial Intelligence," *Annals of Internal Medicine,* vol. 4, nr. 8, 2025.

[9] The Guardian, „Man develops rare condition after ChatGPT query over stopping eating salt,” 12 08 2025. [Interactiv]. Available: https://www.theguardian.com/technology/2025/aug/12/us-man-bromism-salt-diet-chatgpt-openai-health-information. [Accesat 03 11 2025].

[10] BBC, „ChatGPT shares data on how many users exhibit psychosis or suicidal thoughts,” 28 10 2025. [Interactiv]. Available: https://www.bbc.com/news/articles/c5yd90g0q43o. [Accesat 03 11 2025].

[11] Futurism, „OpenAI Says It’s Scanning Users’ ChatGPT Conversations and Reporting Content to the Police,” 27 08 2025. [Interactiv]. Available: https://futurism.com/openai-scanning-conversations-police. [Accesat 30 10 2025].

[12] OpenAI, „How people are using ChatGPT,” 15 09 2025. [Interactiv]. Available: https://openai.com/index/how-people-are-using-chatgpt/. [Accesat 30 10 2025].

[13] A. Chatterji, T. Cunningham, D. Deming, Z. Hitzig, C. Ong, C. Y. Shan și K. Wadman, „How People Use ChatGPT,” National Bureau of Economic Research, Working Papers, 2025.

[14] H. Morrin, L. Nicholls, M. Levin, J. Yiend, U. Iyengar, F. DelGuidice, S. Bhattacharya, S. Tognin, J. MacCabe, R. Twumasi, B. Alderson-Day și T. Pollak, „Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it),” *PsyArXiv,* 2025.

[15] The Guardian, „OpenAI stops ChatGPT from telling people to break up with partners,” 05 08 2025. [Interactiv]. Available: https://www.theguardian.com/technology/2025/aug/05/chatgpt-breakups-changes-open-ai. [Accesat 03 11 2025].

[16] OpenAI, „Safety at every step,” 29 09 2025. [Interactiv]. Available: https://openai.com/safety/. [Accesat 30 10 2025].

[17] L. Osler, „Hallucinating with AI: AI Psychosis as Distributed Delusions,” arXiv, Cornell University, 2025.

[18] Medium, „AlphaEvolve,” 25 07 2025. [Interactiv]. Available: https://cbarkinozer.medium.com/alphaevolve-419242a0a4b4. [Accesat 30 10 2025].

[19] Medium, „The Darwin Gödel Machine: Open-Ended Improvement via Recursive Code Mutation and Empirical Fitness,” 24 03 2025. [Interactiv]. Available: https://medium.com/@adnanmasood/the-darwin-g%C3%B6del-machine-open-ended-improvement-via-recursive-code-mutation-and-empirical-fitness-a777681d73e4. [Accesat 30 10 2025].

[20] Sakana.ai, „The Darwin Gödel Machine: AI that improves itself by rewriting its own code,” 30 05 2025. [Interactiv]. Available: https://sakana.ai/dgm/. [Accesat 30 10 2025].

[21] D. Tapscott, Grown Up Digital: How the Net Generation is Changing Your World, McGraw Hill, 2009.

[22] A.-M. Tirziu și C. Vrabie, „NET Generation. Thinking outside the box by using online learning methods,” *MPRA,* 2016.

[23] C. Vrabie, „Book Review: 'Grown Up Digital: How the Net Generation is Changing Your World' by Don Tapscott,” *HOLISTICA Journal of Business and Public Administration,* vol. 6, nr. 1, 2015.

[24] D. Shapiro, „AI Psychosis. I should have seen this coming,” Substack, 14 08 2025. [Interactiv]. Available: https://daveshap.substack.com/p/ai-psychosis. [Accesat 10 11 2025].