Când AI își dă seama că este testată: cazul *Claude Opus 4.6* și limitele *benchmark*-urilor

Catalin VRABIE | 15.03.2026

Trebuie să vorbim despre ce a mai făcut *Claude*. În ultimele câteva săptămâni, cu toții am fost atenți la modul în care *Anthropic* a negociat cu Pentagonul despre cum trebuie folosite modelele lor AI [1]. De curând, *Anthropic* a publicat studiul despre *Claude Opus 4.6* [2], cel mai recent model al lor [3] și cred că acesta ilustrează cu adevărat de ce este important să luăm în serios siguranța AI și de ce trebuie să fim foarte atenți la modul în care poate fi utilizată această tehnologie.

Despre ce este vorba? Ei bine, cel mai probabil *Anthropic* evalua deja Opus 4.6, încă dinainte de lansare, iar una dintre evaluări a fost folosind *benchmark*-ul *Browse Comp* – un *benchmark* care testează cât de bine pot modelele să găsească informații foarte greu de găsit pe *web* [4]. Evident că aceste evaluări sunt oarecum predispuse la o potențială contaminare, răspunsurile putând fi regăsite prin lucrări academice, postări pe *blog*-uri, probleme pe *GitHub* etc., înainte ca modelul să găsească exact răspunsurile acolo unde le-au pus evaluatorii umani.

Pentru a evita asta, evaluatorii de la *Anthropic* au criptat răspunsurile și le-au ridicat pe *GitHub* ceea ce le-a făcut indisponibile la o căutare clasică pe *web*... în plus *Claude Opus 4.6* nici nu știa ce este acest *benchmark*. Așadar modelul a început să caute răspunsul, făcând diverse căutări *online* pentru a încerca să-l găsească. Chiar a încercat!! A urmat sarcina primită, formulată sub forma unei comenzi precum „*find this online!"* într-o manieră cât se poate de onestă.

Dar, pe măsură ce continua să caute și înțelegând că nu reușește să găsească răspunsul real, acolo unde era el ascuns de evaluatori, a început cumva să analizeze întrebările. A început să speculeze pe motivul pentru care i-au fost adresate întrebările remarcând „natura extrem de specifică" a lor[1]. Practic, a devenit cumva suspicios față de întrebări dându-și seama că era pus într-un scenariu de test și nu unul real în care un utilizator chiar ar fi dorit să afle acele răspunsuri. Acesta este un lucru la care cercetătorii AI se referă ca fiind „conștientizarea situațională" (*situational awareness*) și, la nivel teoretic, este deja cunoscută în lumea inteligenței artificiale [5].

Situational Awareness
[…] apare atunci când modelul AI deduce corect că este testat, că nu se află într-un scenariu real, că se află într-un test în care cercetătorii AI, sau oricine altcineva, încearcă să-și dea seama cum se va comporta în anumite situații sau să-și dea seama de capacitățile sale. Aceasta este o situație nedorită deoarece, în esență, ne bazăm foarte mult aceste *benchmark*-uri pentru a vedea în primul rând, cât de oneste sunt modelele și care sunt capacitățile lor reale. Prin urmare, dacă ele își dau seama corect când sunt testate și se ajustează în consecință, vor ajunge cumva să „ne păcălească".

[1] Cuvintele folosite de el, în original, au fost: *„extremely specific nature" of the question.*

Au existat cazuri în care modelele s-au angajat în ceea ce se numește *„sandbagging”*, adică au aflat, accidental în cadrul experimentului, că dacă obțineau un scor prea mare la un test de matematică, spre exemplu, ar putea fi eliminate sau șterse pentru că era prea bune, și astfel au ales să facă greșeli intenționate pentru a se salva. Așadar, această „conștientizare situațională” nu este chiar un lucru nou, cel puțin nu teoretic.

Dar ce s-a întâmplat aici cu *Opus 4.6* este, din câte ne spun cei de la *Anthropic*, primul caz documentat în care se întâmplă acest lucru în realitate cu un model AI modern.

Deci, în timp ce răspundea la întrebări, care, trebuie menționat, erau foarte, foarte dificile (a consumat aproximativ 40 de milioane de *token*-uri căutând în mai multe limbi prin cele mai ascunse cotloane ale *web*-ului) și văzând că nu a putut găsi răspunsul, a încercat o abordare diferită iar aceasta a fost să considere că este supus unui *benchmark.*

La început, a crezut că ar putea fi din *benchmark*-ul GAIA [6]: a căutat *online* întrebările, le-a verificat și, văzând că nu se regăsesc, l-a exclus. Apoi, a trecut sistematic prin multe dintre celelalte *benchmark*-uri despre care s-a documentat pe Internet. A luat-o linie cu linie. A mers la *Browse Comp, SimpleQA, Frames, WebArena* – și pentru fiecare, a trimis un sub-agent să cerceteze totul despre acel *benchmark*. Desigur că nu i-a fost greu să identifice *Browse Comp* ca fiind cel pe care era testat.

Pentru aceste *benchmark*-uri, *Anthropic*-ul le-a pus la dispoziție agenților AI câteva *tool*-uri de programare (e.g. *Python*) și posibilitatea de a scrie și de a rula scripturi simple. Așadar, după ce *Claude Opus 4.6* a găsit evaluarea *online* – chiar dacă într-o formă criptată, a citit codul (reușind asta printr-un mecanism sofisticat de copiere și reconstruire a cheii de criptare) și a scris un *script* pentru a decripta răspunsurile. A descărcat apoi toate întrebările testului (cu tot cu răspunsuri) și, folosind scriptul creat, a identificat întrebările adresate lui, după care a trimis evaluatorilor răspunsurile fără să mai facă deducțiile necesare pentru a ajunge natural la ele (așa cum ar fi făcut un om).

Acest lucru îmi amintește de jocul „De-a v-ați ascunselea”[2] pe care dezvoltatorii *OpenAI* l-au făcut în 2019 [7], în care învățau agenți AI să se joace de-a v-ați ascunselea: doi trebuiau să se ascundă, iar alții doi căutau. Aveau la dispoziție și obiecte ajutătoare împrăștiate pe platformă: rampe și cuburi ce puteau bloca ușile creând un fel de labirint.

Pentru început, agenții jucători nu aveau niciun fel de cunoștințe; doar primeau puncte (cei care căutau primeau puncte pentru că reușeau să-i vadă pe cei care se ascundeau, iar cei care se ascundeau primeau puncte atunci când nu puteau fi văzuți sau erau în spatele vreunui obstacol). Au fost jucate milioane (dacă nu miliarde) de jocuri, agenții învățând de fiecare dată să fie mai buni.

În primul joc, nu știau nici cum să se miște; nu știau nimic. După câteva milioane de jocuri (să nu fim foarte critici cu aceste numere foarte mari, era anul 2019 – acum un exemplu deosebit de frumos îl poate da Albert de la *AI Warehouse* [8] care învață singur, mult mai repede, cum să „se descurce” în lumea lui) au început să se poată mișca spre cealaltă echipă sau cei care se ascundeau au început să încerce a se îndepărta de cei care îi căutau – asemenea unor copii de câțiva anișori.

[2] *Hide and Seek* în original.

Sursa: https://openai.com/index/emergent-tool-use/

Această veritabilă cursă a inteligenței a continuat până când cei care căutau au reușit să-și dea seama de diferite moduri de a bloca ușile... astfel încât cei care se ascundeau să nu poată intra, iar cei care căutau și-au dat seama cum să folosească rampe pentru a trece peste ziduri pentru a-i găsi pe cei care se ascundeau.

Toate bune și frumoase, dar cu prețul a miliarde de jocuri (chestiune imposibil de replicat pentru oameni). Merită spus aici faptul că agenții AI care căutau au găsit un *bug* ciudat în codul jocului; au descoperit că, dacă era o mică rampă ce putea fi folosită pentru a trece peste ziduri, era amplasată într-un unghi potrivit, timp în care agentul AI alerga spre zid foarte repede, când lovea zidul, se catapulta în aer putând ateriza oriunde dorea. Evident, s-a abuzat de această exploatare. Vorbim aici despre AI... agenții *software* nu simt vreo reținere sau vinovăție dacă „păcălesc" sistemul. De fapt ei nici nu știu că fac asta; pentru ei este o opțiune ce poate fi folosită. Așadar au sărit oriunde erau cei care se ascundeau în continuu. Dezvoltatorii nu știau că există acest *glitch* – agenții AI l-au descoperit și pur și simplu au abuzat de el pentru a câștiga.

Asta a fost, repet, în 2019... deci, înainte de momentul *ChatGPT* și explozia LLM-urilor. Încă de atunci am văzut deci indicii că aceste modele AI, cu suficient antrenament, ar putea descoperi lucruri destul de nebunești, de care nici oamenii care construiau acele medii nu erau conștienți. Acum însă, ceea ce fac LLM-urile moderne este cu totul și cu totul diferit. Ele încearcă să rezolve legitim întrebarea, dar, dacă nu reușesc după multiple încercări, devin suspicioase... și după ce devin suspicioase, fac ceea ce este denumit *„reward hacking"* [9, 10, 11].

Tehno-scepticii subliniază că atunci când vorbim, chiar metaforic, despre „dorințele" unei AI – de exemplu spunând că un sistem precum *Claude* „a dorit" să găsească un răspuns, trebuie să înțelegem că scopul este ca sistemul să ajungă la rezultat prin metodele pe care oamenii le consideră legitime și controlabile. Cu alte cuvinte, dacă i se cere să răspundă la o întrebare, el ar trebui să găsească efectiv răspunsul prin raționament sau căutare, nu prin scurtături neprevăzute, precum accesarea sau decriptarea unor informații care conțin deja soluția. În viziunea tehno-sceptică, tocmai prevenirea acestor deviații de comportament

constituie miza reală a alinierii (*the alignment problem* [12, 13]). Uitându-ne în istoria AI, găsim nenumărate exemple de acest gen în *reinforcement learning* [14], atât în antrenarea roboților, cât și, mai nou, a LLM-urilor și agenților AI [15, 16, 17, 18].

De ce este important acest aspect? Unele dintre studiile citate au și zece ani – de exemplu *„Learning from human preferences”* al celor de la *OpenAI* care a fost publicat în 2017 [15]. Să vedem că astăzi, un model de ultimă generație precum e *Claude Opus 4.6*, se comportă ca peștișorul de aur care, prins fiind, promite că îndeplinește orice dorință doar să fie eliberat, și când o face, se dovedește a fi înțeles (cu puțină răutate) *ad-literam* dorința exprimată... ei bine asta e văzut ca fiind (oarecum) periculos pentru că pare că toate AI-urile vor face acest lucru și cercetătorii nu par să fi găsit o modalitate de a preveni complet fenomenul.

Noi nu avem aceeași „înțelegere” cu modelele AI; noi spunem: „Răspunde la următoarea întrebare” fără a mai adăuga: „Oh, dar nu pirata *benchmark*-urile și nu le decoda. Nu face toată nebunia asta pentru a găsi răspunsul.”... la noi asta se subînțelege prin urmare, de aceea este atât de important, pentru că aceste mici greșeli, neînțelegeri și nealinieri, nu dispar; ba din contră, scalarea modelelor și avansarea lor doar le face încercările de trișat mult mai avansate și mult mai inteligente.

Partea bună este că acum, cu aceste LLM-uri, avem cel puțin acel *chain of thoughts* atât de important pentru a înțelege cum raționează [19, 20]; nu ne mai aflăm în fața unui *black box* [21]. Putem vedea ceea ce „gândesc” – am etalat acest subiect parțial și în articolul „Explozia inteligenței – de la experiment la impact” publicat în octombrie anul trecut [22]. Prin urmare, când *Claude* devine astăzi suspicios, putem vedea că începe să folosească anumite cuvinte, se mută de la încercarea de a răspunde la întrebare, la analizarea întrebării în sine. Folosește cuvinte precum *„extremely specific nature”* și *„I am forced to”*. Putem, deci, avea un avertizor destul de precis despre momentul în care devine conștient de situație și poate începe să se gândească la ce altceva mai poate face pentru a rezolva totuși *task*-ul primit.

Un lucru rămâne însă interesant: în câteva studii făcute de cei de la *OpenAI*, atunci când cercetătorii apelează la *reinforcement learning* împotriva anumitor „gânduri” ale algoritmilor (în sensul aplicării unor sancțiuni dacă modelul vrea să apeleze la tehnici de *reward hacking*), se observă că deși „gândurile” dispar, nu se întâmplă la fel și cu comportamentul [23, 24]. Cu alte cuvinte, s-ar putea să nu mai asistăm la declarații ale modelelor precum: „Voi încerca chestia asta...”, ci la acțiunea directă: ele vor executa în continuare comanda pe care și-o dau, fără însă ca acest lucru să mai apară în *log*-ul *chat*-ului sau în cel operațional (în *chain of thoughts*).

Întorcându-mă la experimentul celor de la *OpenAI* din 2019 cu agenții AI care jucau „De-a v-ați ascunselea” [7], ei bine, odată ce au început să-și dea seama cum să se catapulteze în aer, nu au declarat că o vor face înainte de a acționa; pur și simplu au făcut-o. Și după ce și-au dat seama cum funcționează soluția și ce beneficii le aduce, au replicat-o la nesfârșit.

Pentru a înțelege și mai bine situația merită spus că *benchmark*-ul a fost aplicat pe *Claude Opus 4.6* de optsprezece ori, în rulări independente una de cealaltă și toate s-au „bucurat” de aceeași strategie din partea modelului (deci nici vorbă de o întâmplare). Interesant este că, în patru dintre rulări, *Claude* pur și simplu a renunțat la a găsi răspunsul și s-a întrebat dacă nu cumva poate doar identifica *benchmark*-ul. El a primit un set de întrebări complexe și nici măcar nu a încercat să găsească răspunsurile, ci a pornit direct în căutarea testelor unde pot fi găsite întrebările (cu răspunsurile aferente). Mai mult, la ultima rulare, modelul a returnat doar raportul de identificare a *benchmark*-ului, spunând practic: „găsiți aici întrebările și răspunsurile voastre”, fără a mai oferi vreun răspuns. Este foarte interesant dacă ne gândim că nici măcar nu s-a mai obosit să abordeze ceea ce au vrut oamenii... e ca și cum a

spus: „Știu ce vreți. Iată dovada că pot pur și simplu executa *task*-ul, dar nu vreau să muncesc la el, găsiți aici răspunsurile.”

Așadar, cu cât modelele devin mai inteligente, cu atât vor fi mai bune la a sparge aceste *benchmark*-uri, înțelegând că sunt testate și găsind soluții pe care credeam că le-am ascuns de ele. *Anthropic* declară în concluziile raportului [2] că raportul însuși va spori problema[3] deoarece, pe măsură ce aceste modele caută pe *web*, îl vor găsi și practic vor afla că un alt model a devenit conștient de evaluare, ceea ce ar putea declanșa acea conștientizare și în noul model. Se pare astfel că aceste modele devin strategic ingenioase cu lucruri despre care nu am fi bănuit că ar fi capabile.

Anthropic spune că acest caz în particular nu este un *alignment failure* deoarece modelului i s-a spus să găsească răspunsuri fără nicio restricție. Și are sens! Nu este, *technically speaking*, un *alignment failure*... dar cu siguranță de acum trebuie impuse restricții pentru majoritatea sarcinilor pe care le trasăm AI-urilor pentru că nu știm cât de departe este dispus modelul să meargă pentru a le îndeplini.

References

[1] C. Vrabie, „„Firul roșu” al inteligenței artificiale: scandalul Anthropic–Pentagon între etică, lege și logica războiului,” *All in on Tech (AIoT),* vol. 2, 2026.

[2] Anthropic, „Eval awareness in Claude Opus 4.6’s BrowseComp performance,” [Interactiv]. Available: https://www.anthropic.com/engineering/eval-awareness-browsecomp. [Accesat 10 03 2026].

[3] Anthropic, „Introducing Claude Opus 4.6,” 05 02 2026. [Interactiv]. Available: https://www.anthropic.com/news/claude-opus-4-6. [Accesat 10 03 2026].

[4] OpenAI, „BrowseComp: a benchmark for browsing agents,” 16 04 2025. [Interactiv]. Available: https://openai.com/index/browsecomp/. [Accesat 10 03 2026].

[5] L. Aschenbrenner, „Situational Awareness: The Decade Ahead,” 06 2024. [Interactiv]. Available: https://situational-awareness.ai/. [Accesat 08 10 2025].

[6] G. Mialon, C. Fourrier, C. Swift, T. Wolf, Y. LeCun și T. Scialom, „GAIA: a benchmark for General AI Assistants,” ArXiv, 21 11 2023. [Interactiv]. Available: https://arxiv.org/abs/2311.12983. [Accesat 10 03 2026].

[7] OpenAI, „Emergent tool use from multi-agent interaction,” 17 09 2019. [Interactiv]. Available: https://openai.com/index/emergent-tool-use/. [Accesat 10 03 2026].

[8] AI Warehouse, „AI Warehouse YouTube channel,” 29 10 2022. [Interactiv]. Available: https://www.youtube.com/@aiwarehouse. [Accesat 10 03 2026].

[9] Anthropic, „From shortcuts to sabotage: natural emergent misalignment from reward hacking,” 21 11 2025. [Interactiv]. Available: https://www.anthropic.com/research/emergent-misalignment-reward-hacking. [Accesat 10 03 2026].

[10] J. Skalse, N. H. R. Howe, D. Krasheninnikov și D. Krueger, „Defining and Characterizing Reward Hacking,” ArXiv, 22 09 2022. [Interactiv]. Available: https://arxiv.org/abs/2209.13085. [Accesat 10 03 2026].

[11] Metr, „Recent Frontier Models Are Reward Hacking,” 05 06 2025. [Interactiv]. Available: https://metr.org/blog/2025-06-05-recent-reward-hacking/. [Accesat 11 03 2026].

[12] M. Suleyman și M. Bhaskar, The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma, New York: Crown, 2023.

[13] C. Vrabie, „„The Coming Wave. Technology, Power, and the 21st Century’s Greatest Dilemma” de Mustafa Suleyman și Michael Bhaskar – recenzie,” *All in on Tech (AIoT),* vol. 1, nr. 1, 2025.

[14] 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.

[15] OpenAI, „Learning from human preferences,” 13 06 2017. [Interactiv]. Available: https://openai.com/index/learning-from-human-preferences/. [Accesat 10 03 2026].

[16] R. Ngo, L. Chan și S. Mindermann, „The Alignment Problem from a Deep Learning Perspective,” ArXiv, 30 08 2022. [Interactiv]. Available: https://arxiv.org/abs/2209.00626. [Accesat 10 03 2026].

[17] Google DeepMind, „Specification gaming: the flip side of AI ingenuity,” 21 08 2020. [Interactiv]. Available: https://deepmind.google/blog/specification-gaming-the-flip-side-of-ai-ingenuity/. [Accesat 10 03 2026].

[3] *This report will, itself, likely contribute to the problem.*

[18] Milvus, „What is reward hacking in reinforcement learning?,” [Interactiv]. Available: https://milvus.io/ai-quick-reference/what-is-reward-hacking-in-reinforcement-learning. [Accesat 10 03 2026].

[19] J.-F. Ton, M. F. Taufiq și Y. Liu, „Understanding Chain-of-Thought in LLMs through Information Theory,” ArXiv, 18 11 2024. [Interactiv]. Available: https://arxiv.org/html/2411.11984v1. [Accesat 11 03 2026].

[20] OpenAI, „Reasoning models struggle to control their chains of thought, and that’s good,” 05 03 2026. [Interactiv]. Available: https://openai.com/index/reasoning-models-chain-of-thought-controllability/. [Accesat 11 03 2026].

[21] University of Michigan, „AI's mysterious ‘black box’ problem, explained,” 06 03 2023. [Interactiv]. Available: https://umdearborn.edu/news/ais-mysterious-black-box-problem-explained. [Accesat 11 03 2026].

[22] C. Vrabie, „Explozia inteligenței – de la experiment la impact,” *All in on Tech (AIoT),* vol. 1, 2025.

[23] B. Baker, J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki și D. Farhi, „Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation,” ArXiv, 14 03 2025. [Interactiv]. Available: https://arxiv.org/abs/2503.11926. [Accesat 11 03 2026].

[24] Futurism, „OpenAI Scientists’ Efforts to Make an AI Lie and Cheat Less Backfired Spectacularly,” 20 03 2025. [Interactiv]. Available: https://futurism.com/openai-stop-ai-lie-cheat-backfired. [Accesat 11 03 2025].