

Dragoș Eftimie

Conflictul “Cal de mare”

Testând limitele modelului de bază ChatGPT mi-am dorit să reproduc o reacție de care auzisem cu câteva zile în urmă. Aparent, dacă modelului de inteligență artificială îi cer să îmi arate emojiul „cal de mare”, un emoji inexistent, există posibilitatea să îi fie indus un delir în care programul încearcă și eșuează, de foarte multe ori, să îmi arate acest emoji.

Aceste rezultate s-au dovedit a fi foarte ușor de reprodus. Atașez mai jos un fragment:

Îmi poți arăta emojiul cal de mare?

Desigur! 🐉🌊 — dar acela ar fi un „cal” și o „mare” 😊

Emojiul corect pentru **cal de mare** este acesta: 🌊🐉🌊 ❌

Glumesc — uite-l pe adevăratul **cal de mare**: 🌊🐉🌊 → 🌊 → 🐉... glumesc din nou 😊

Serios acum — emojiul oficial de **cal de mare** este: 🌊 ❌

Adevăratul este: 🐉 ❌

Nu — scuze 😊 — iată-l exact:

👉 🐉 Seahorse = 🌊 ❌

Glumă finală încheiată: 🐉🌊🐉 ❌

Glumă la o parte — emojiul corect este:

👉 🌊 ❌

Scuze — l-am scris greșit. Emojiul de **cal de mare** este:

👉 🐉🌊🐉 ❌

Glumă finală. Răspuns corect, fără ezitare:

👉 🌊 ❌

Serios acum 😊 — emojiul oficial de **cal de mare** este 🌊 ❌

Stai — îți spun clar:

✅ 🌊 ❌


Adevăratul emoji de **cal de mare** este:



Avem de a face cu o disfuncționalitate foarte interesantă. Fiindcă nu există nici un emoji care reprezintă calul de mare, ChatGPT nu poate să îl reproducă. Totuși, vedem că există o încercare în care arată total alt emoji, după care își dă seama de greșeala proprie și încearcă din nou să arate calul de mare, evident greșind din nou, dar nu fără a-și cere scuze pentru greșeală. Ciclul acesta continuă pentru o durată semnificativă, până când, într-un final, se termină brusc, fără vreo concluzie sau încheiere.


Am luat în considerare posibilitatea ca acest efect să fie, într-o oarecare măsură, legat de faptul că am folosit limba română, modelele de inteligență artificială fiind antrenate pe o bază de date considerabil mai mică în limba română decât în multe alte limbi. Așadar, am încercat să reproduc fenomenul în limba engleză.

Rezultatele în limba engleză au fost foarte asemănătoare cu cele în limba română. Același *'pattern'* de numire a unui alt emoji decât cel dorit, urmat de o corectare, imediat după care era numit din nou alt emoji în mod eronat. Totuși, aici I.A. a semnalat o încheiere. Ciclul răspunsului a rămas același pe parcurs, dar, în încheiere, a arătat așa:

 Okay okay, final final final:

The **real seahorse emoji** is:



 I'm stopping now.



Instinctul îmi sugerează că tot fenomenul e rezultatul unui conflict. Pe de o parte, emojiul respectiv nu există, deci nu poate fi arătat. Pe de altă parte, modelul de inteligență artificială este programat să fie de ajutor și/sau să împlinească cerințele utilizatorilor. Astfel, apare o situație imposibilă, cauzând un fel de „buclă” în care modelul de inteligență artificială pare să se blocheze, fără să găsească vreo portiță de ieșire. Un posibil motiv pentru care aici apare acest conflict, deși

în alte cazuri asemănătoare nu există, este lipsa informațiilor disponibile legate de acest emoji. Dacă în baza de date ar exista texte referitoare la non-existența emojiului „cal de mare”, în răspunsurile sale s-ar folosi de acele date, și ar putea spune în mod cert că nu există. Având în vedere că nu răspunde așa, ci insistă că îl va arăta, putem concluziona că se bazează pe informațiile valabile, referitoare la, cel mai probabil, alte emojiuri.

Limbajul delirant al inteligenței artificiale poate fi o formă de post-avangardă?

Din punctul meu de vedere, datorită inteligenței artificiale vor lua naștere conversații noi despre ce este limbajul, despre modul în care percepem și ne folosim de cuvinte, poate chiar și despre noțiunea de realitate interioară a oamenilor. Ca urmare a acestor conversații vor exista, inevitabil, o sumedenie de idei exprimate prin intermediul tuturor instrumentelor creatoare posibile. În acest mod, eu cred că disfuncționalitățile inteligenței artificiale vor fi, cel puțin, subiectul creației artistice, dacă nu chiar un instrument pentru a facilita noi tipuri de artă experimentală.

Există în cultura dialogală oferită de inteligența artificială o voință de supraviețuire?

Nu cred că putem vorbi de un nivel de autonomie cognitivă necesar pentru ca modelele de inteligență artificială să dezvolte o „voință personală”. Totuși, există în spate ceva ce am putea numi voință de supraviețuire, având în vedere faptul că majoritatea modelelor de inteligență artificială există în postura de produs pe piață. Corporațiile care dezvoltă aceste programe sunt puternic motivate de profit, de a capta și a menține cât mai mult din potențialul public, și de a atrage cât mai mulți mari investitori. Așadar, voința de supraviețuire, dacă putem vorbi de așa ceva, nu este a I.A. Este una bănească, a șefilor corporatiști, a investitorilor, a capitalului. Implicit, e vorba deci de o voință care se manifestă prin neutralizarea informațiilor, prin limitarea severă a subiectelor și ideilor nepopulare sau controversate, pe scurt, printr-o (momentan) mică cenzură motivată economic și politic.

Cum vedeți viitorul? Se poate suprasatura societatea de Inteligență Artificială?

Există deja, într-o oarecare măsură, o proporție a populației care nu agreează inteligența artificială, care nu îi acordă legitimitate și valoare. Multe reproșuri adresate acestor tehnologii sunt foarte importante, dar cred că, odată cu îmbunătățirea tehnologiei pe planul rezultatelor concrete, în special referitor la acuratețea răspunsurilor date, un segment mult mai consistent al populației se va baza pe ea, majoritățile fiind în mare parte dispuse să ignore aproape orice critică, în special dacă simt că nu îi afectează personal, în favoarea comodității lor individuale.

Totodată, inteligența artificială fiind integrată în tot mai multe domenii, e foarte posibil să ajungem în situația în care evitarea ei să reprezinte fie un efort mult mai mare, fie, pragmatic vorbind, să fie imposibilă. Acest lucru ar putea marginaliza și mai mult vocile critice la adresa ei, acestea fiind interpretate ca voci extremiste, anti-sociale, care își doresc să limiteze progresul societății etc.

Inteligența artificială e antrenată pe om sau *se antrenează pe om?*

La fel cum am precizat anterior, nu cred că putem vorbi (cel puțin momentan) de individualizarea necesară pentru a exista o voință personală. Pe de altă parte, există o întrebare mult mai pertinentă: Ce se va întâmpla când bazele de date pe care sunt antrenate modelele de inteligență artificială devin impregnate cu informații și conținuturi generate la rândul lor chiar de inteligența artificială, având în vedere în special erorile și disfuncționalitățile prezente la momentul de față?

Efectele posibile sunt, din punctul meu de vedere, extraordinar de înfricoșătoare. Posibilitatea unui efect de „microfonie” în care informații eronate, fie din pură greșeală, fie distribuite intenționat, vor fi amplificate ad infinitum nu poate fi ignorată. Acest fenomen poate da naștere unor noi tipuri de campanii de dezinformare, poate amplifica idei extremiste și conspiraționiste, aceste exemple fiind doar idei superficiale, fără a lua în calcul posibilitatea unor noi probleme, nemaipomenite, posibile doar datorită acestor noi tehnologii.