Superinteligența – una dintre cele mai mari provocări tehnice a momentului

Catalin VRABIE / 20.01.2025

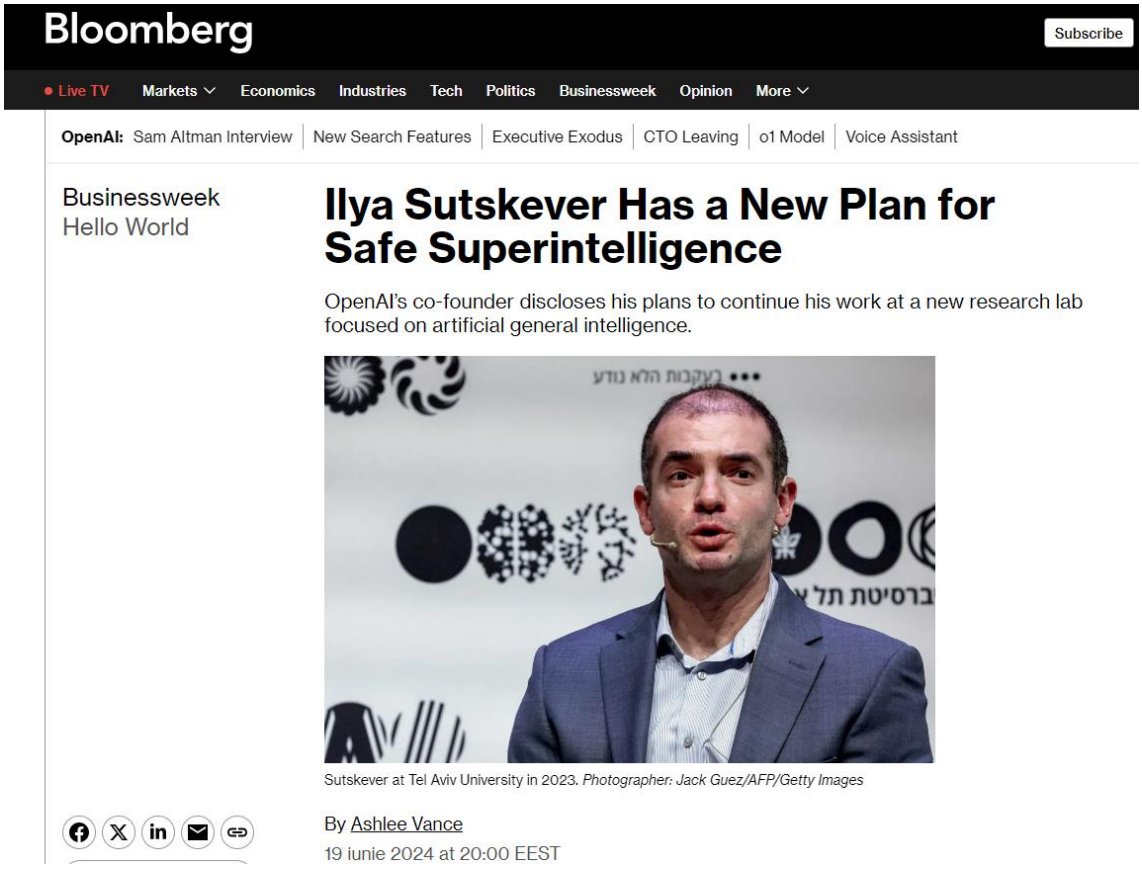
„Cu siguranță vom putea crea entități complet autonome, cu obiective proprii, și va fi foarte important, mai ales pe măsură ce acestea devin mult mai inteligente decât oamenii, ca obiectivele lor să fie aliniate cu ale noastre.” [1] Aceasta declarație aparține uneia dintre cele mai misterioase figuri din domeniul Inteligenței Artificiale (AI), Ilya Sutskever. Ilya este un informatician reputat, direct implicat în dezvoltarea OpenAI și ChatGPT, susținut și apreciat pentru cercetările sale, printre alții, de Sam Altman, Elon Musk, Jensen Huang și Geoffrey Hinton.

CEO-ul NVIDIA, Jensen Huang în dialog cu Ilya Sutskever la GTS 2023

Co-fondator și fost director științific la OpenAI (pentru cei care nu știu, Ilya fost unul dintre membrii Consiliului de administrație care au decis demiterea lui Sam Altman din poziția de CEO, invocând lipsa de încredere în capacitatea acestuia de a continua să conducă organizația. A revenit asupra deciziei dar, după reinstituirea lui Altman, și-a dat demisia din OpenAI[1]), Ilya este constant în centrul atenției – memele asociate lui sunt, de fapt, în bună parte, meme ale AI-ului (*Feel the AGI* este, de pildă, un slogan pe care obișnuia să-l tot folosească în birourile OpenAI).

[1] În ceea ce privește relația cu Sam Altman, Ilya declară că a rămas bună, deși întreaga experiență a ultimelor luni la OpenAI a fost „stranie” [3].

Apariţiile sporadice şi declaraţiile vagi din ultimul timp au trezit curiozitatea multor experţi AI, care au inceput să (se) întrebe unde este Ilya şi căror obiective i se dedică. Ei bine, ştim acum că în 2024 Ilya a fost implicat nu în construirea AGI-ului (Artificial General Intelligence[2]) - un subiect deja prăfuit pentru un om de calibrul său, ci a ASI-ului (Artificial Super Intelligence), pe care o numeşte SSI (Safe Super Intelligence), „cea mai importantă problemă tehnică a vremurilor noastre" [2] şi a pus bazele primului laborator *straight-shot* pentru SSI (voi reveni asupra acestei idei).



[2] Pentru detalii despre ce înseamnă asta prin comparatie cu termenul, oarecum classic, de AI, vă invit să parcurgeţi paginile volumului: „AI: de la idee la implementare. Traseul sinuos al Inteligenţei Artificiale către maturitate" [13].

În vara lui 2024, Ilya Sutskever declara pentru Bloomberg [3] că lansează[3] o companie nouă, SSI.INC, unde, alături de o echipă mică, dar extrem de talentată [4] „va urmări *Safe Superintelligence* într-un mod direct, cu un singur obiectiv și un singur produs". Ca multe alte *start-up*-uri din Silicon Valley, SSI.INC a reușit în cele două luni de la înființare [5] să atragă un capital uriaș – nu mai puțin de un miliard de dolari, fiind astăzi evaluată neoficial la cinci miliarde USD [5]).

Pe *site*-ul companiei scrie clar: „Superinteligența este la îndemână." Pentru a oferi puțin context, trebuie spus că în ultimele luni, mulți experți au sugerat că AI-ul și în special *deep learning*-ul își vor încetini dezvoltarea, urmând să ajungă cât de curând la o stagnare; Yann LeCun explică acest lucru într-un interviu oferit lui Craig Smith de la *Eye on AI* invocând faptul că LLMs (Large Language Models) nu reprezintă calea spre AGI [6]. Ilya, în schimb, în septembrie 2023 infirma tendința, declarând că iarna AI[4] nu va avea loc, iar „AGI și ASI sunt cu siguranță obiective posibil de atins în decursul vieții noastre" [7].

De altfel, siguranța lui Ilya se vede și astăzi, în pagina de Internet a companiei nou înființate, unde cuvântul „sigur" este folosit frecvent în legătură cu progresul AI. Cu birouri în *Palo Alto* și *Tel Aviv*, unde fondatorii (Ilya, Daniel Gross și Daniel Levy) au rădăcini adânci și, după propriile declarații, dețin și capacitatea de a recruta talente tehnice de top[5], SSI.INC se bazează deja pe ingineri și cercetători de vârf. Conducerea sa este la rându-i de excepție. Daniel Gross, inginer și investitor, co-fondator al unui motor de căutare Q&A achiziționat de Apple în 2013, a condus proiecte AI la Apple, investind în startup-uri mari precum *Instacart*, *Coinbase* și *GitHub*. A fost partener la *Y-Combinator* și a inițiat programul de AI al acestui hub [8, 9]. La rândul său, Daniel Levy, doctor al Universității *Stanford* , a lucrat pentru *OpenAI*, *Google Brain* și *Facebook* și confirmă viziunea de leadership a lui Ilya: (avem nevoie de, n.a.) „o echipă mică, valoroasă în care toți membrii sunt interesați doar de dezvoltarea SSI" [3].

Echipa, investitorii și modelul de afaceri sunt toate aliniate pentru a permite companiei să abordeze simultan siguranța și capacitățile informatice, tratându-le ca pe probleme tehnice a căror rezolvare este posibilă prin inovații inginerești și științifice. Scopul declarat al SSI.INC este de a avansa capacitățile mașinilor inteligente cât mai rapid posibil, în condiții de siguranță pentru a-și putea scala activitatea în liniște [2]. Iar fondurile de care dispune îi garantează independența de presiunile comerciale cu care, de regulă, noile (mici) companii se confruntă.

Ashley Vance, autorul uneia dintre biografiile lui Elon Musk[6], l-a intervievat pe Ilya despre noua sa inițiativă iar acesta și-a exprimat speranța de a-și continua eforturile fără intenția de a concura cu *OpenAI*, *Google* sau *Anthropic* (și abaterile de la scopul inițial ce ar putea apărea odată cu asta) [3]. Compania este în felul ei cumva unică, deoarece primul său produs (și singurul) va fi SSI-ul; nu vor exista demonstrații, nici lansări, nimic, până când obiectivul nu va fi atins. Această abordare ține compania departe de presiunile de a lansa produse sau de a rămâne competitivă pe această piață care în acest moment se află în plină efervescență.

Cu privire la aspectul siguranței, Ilya rămâne destul de vag, dar sugerează că aceasta va fi realizată prin inovații inginerești integrate în sistem și nu ca până acum, prin măsuri aplicate după dezvoltare: „Prin (AI) 'sigură', ne referim la 'siguranță' precum cea nucleară, nu la 'încredere și siguranță'" a declarat el, formulând ceea ce pare a fi o critică subtilă la adresa *OpenAI* care pare că se fundamentează întocmai pe o asemenea abordare [10].

Investitorii în SSI.INC par să susțină proiectul fără a se aștepta profituri rapide. Termenul *straight-shot* SSI, la care am promis că mă voi întoarce, reflectă tocmai acest *focus* – fără câștiguri rapide sau lansări

---

[3] Alături de Daniel Gross și Daniel Levy.

[4] Termen cunoscut în literatura de specialitate ca AI winter – din nou invit cititorii să parcurgă volumul „AI: de la idee la implementare. Traseul sinuos al Inteligenţei Artificiale către maturitate" [13] precum și articolul „Deep Learning. Viitorul inteligenței artificiale și impactul acesteia asupra dezvoltării tehnologiei" [12] pentru a înțelege mai bine conceptul și contextual în care este folosit.

[5] Ilya, deși născut în Rusia, a emigrat la o vârstă fragedă (cinci ani) în Israel, unde și-a început studiile, mutându-se apoi, la vârsta de șaisprezece ani, în *Toronto*, Canada.

[6] Cea din 2015; Există și o biografie redactată de Walter Isaacson în 2023.

incrementale. Nu au API[7]-uri sau modele de abonament și există un oarecare sentiment de „acum ori niciodată” în abordarea companiei.

Toate acestea se întâmplă însă în condițiile în care nu există un consens în lumea științifică și nici în cea a dezvoltatorilor cu privire la fezabilitatea superinteligenței și nici chiar vizavi de drumul ce ar trebui parcurs spre AGI. Mulți se întreabă dacă LLMs au capacitatea de a raționa, inova sau generaliza dincolo de datele lor de antrenament. Totuși, Ilya pare încrezător că superinteligența este *within reach* și unei echipe mici și dedicate, nu numai unor companii cotate la de trilioane de dolari [11] și care se bucură de lansări frecvente de produse. El susține că AI ar trebui să reflecte valorile democrației și libertății, fundamentale societăților dezvoltate, și își imaginează un AI cu scop general, un super centru de date care dezvoltă autonom tehnologie.

Cât de realist este obiectivul lui Sutskever de a construi superinteligența sigură, cu o echipă mică și finanțare limitată rămâne să vedem. Ideea de a nu prezenta produse intermediare și a ținti exclusiv ASI, îl face să pară concentrat și serios. Unii l-ar putea considera prea ambițios, dar reputația de care se bucură sugerează că nu a promis niciodată mai mult decât poate face. Vom vedea curând dacă și cum SSI.INC va lărgi frontierele AI.

References

[1] The Guardian, *Ilya: the AI scientist shaping the world,* 2023.

[2] Safe Superintelligence Inc., „Superintelligence is within reach.,” Safe Superintelligence Inc., 19 06 2024. [Interactiv]. Available: https://ssi.inc/. [Accesat 11 01 2025].

[3] Bloomberg, „Ilya Sutskever Has a New Plan for Safe Superintelligence,” Bloomberg, 19 06 2024. [Interactiv]. Available: https://www.bloomberg.com/news/articles/2024-06-19/openai-co-founder-plans-new-ai-focused-research-lab?leadSource=uverify%20wall&embedded-checkout=true. [Accesat 11 01 2025].

[4] I. Sutskever, X.com, 2024.

[5] Reuters, „Exclusive: OpenAI co-founder Sutskever's new safety-focused AI startup SSI raises $1 billion,” Reuters, 04 09 2024. [Interactiv]. Available: https://www.reuters.com/technology/artificial-intelligence/openai-co-founder-sutskevers-new-safety-focused-ai-startup-ssi-raises-1-billion-2024-09-04/. [Accesat 11 01 2025].

[6] Eye on AI, *Yann LeCun on World Models, AI Threats and Open-Sourcing,* Eye on AI, 2023.

[7] Far.ai, *Ilya Sutskever - Opening Remarks: Confronting the Possibility of AGI,* Far.ai, 2023.

[8] D. Gross, „Daniel Gross,” [Interactiv]. Available: https://dcgross.com/. [Accesat 11 01 2025].

[9] The Information, „The Superintelligence of AI Investor Daniel Gross,” The Information, 28 08 2024. [Interactiv]. Available: https://www.theinformation.com/articles/the-superintelligence-of-ai-investor-daniel-gross. [Accesat 11 01 2025].

[10] OpenAI, „Safety at every step,” OpenAI, 2024. [Interactiv]. Available: https://openai.com/safety/. [Accesat 11 01 2025].

[11] CompaniesMarketcap, „Largest Companies by Marketcap,” CompaniesMarketcap, 13 01 2025. [Interactiv]. Available: https://companiesmarketcap.com/. [Accesat 13 01 2025].

[12] C. Vrabie, „Deep Learning. Viitorul inteligenței artificiale și impactul acesteia asupra dezvoltării tehnologiei,” *Smart Cities International Conference (SCIC) Proceedings,* vol. 10, p. 9–32, 2022.

[13] C. Vrabie, AI de la idee la implementare. Traseul sinuos al Inteligenței Artificiale către maturitate. [AI from idea to implementation. The winding path of Artificial Intelligence to maturity], ISBN 978-606-26-1851-3, Bucharest: Pro Universitaria, 2024.

[7] *Application Programming Interface* - un set de reguli care permit aplicațiilor *software* să comunice și să schimbe date.