Civilizații sintetice: cum agenții AI dezvoltă roluri, norme și forme de autonomie socială

Catalin VRABIE | 23.06.2026

În timp ce citim acest articol, o mie de agenți AI[1] din mult prea bine-cunoscutul joc *Minecraft* lucrează intens la construirea unei adevărate civilizații proprii [1]. Cercetătorii din *Project Sid* i-au introdus într-un mediu AI și i-au lăsat să se descurce, ceea ce a fost suficient pentru ca ei să creeze o întreagă civilizație. Se trezesc, merg la serviciu, bârfesc despre vecinii lor și... se îndrăgostesc [1].

Au organizat alegeri, dezbat nivelul taxelor și fac multe altele care ar fi mai potrivite oamenilor decât personajelor unui joc video... dar, într-adevăr, acest lucru nu este doar despre *Minecraft;* este, mai degrabă, un indiciu asupra unui fenomen mai larg: societăți artificiale populate de agenți autonomi. Potrivit laureaților Premiului Nobel, pionieri ai inteligenței artificiale, Geoffrey Hinton și John J. Hopfield, precum și altor cercetători din industria AI, acesta ar putea deveni curând extrem de important [2, 3, 4].

De la NPC-uri la agenți autonomi

Când ne gândim la personajele AI din jocuri (mă refer aici la NPC - *non-playable character*[2]), probabil că cei mai mulți dintre noi își imaginează personaje simple, care urmează doar câteva reguli de tipul *if-then-else*, însă aceste NPC-uri nu sunt nici autonome și nici foarte interesante; ele putând executa, cel mult, câteva sarcini, și acestea destul de clare și ușor de identificat de către jucătorul uman.

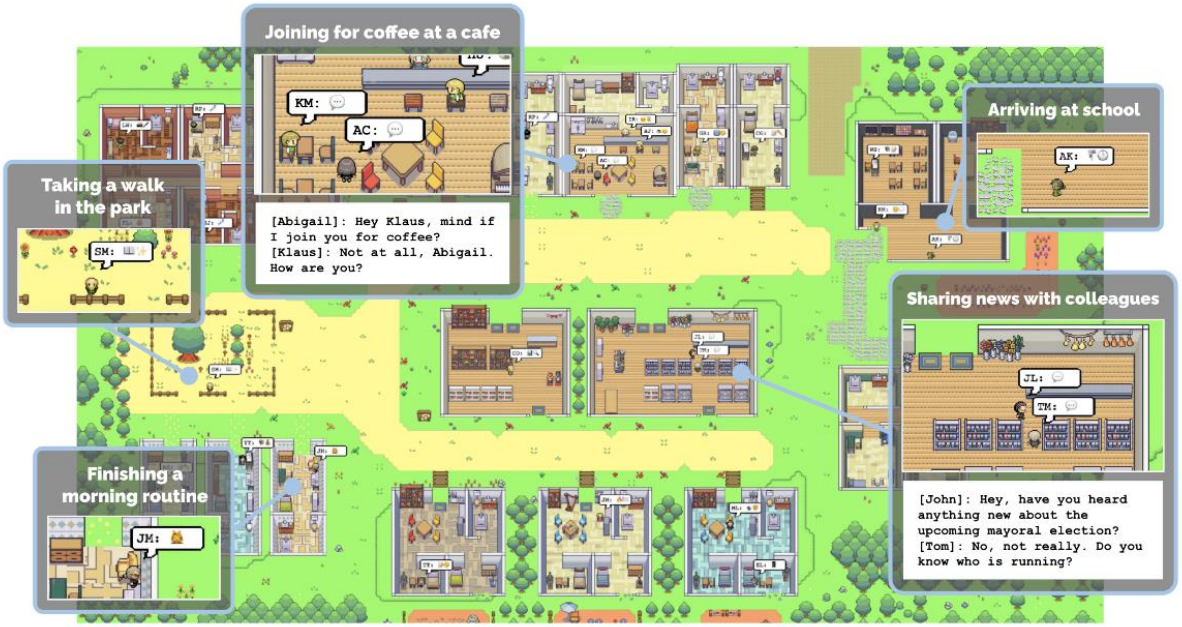
În experimentul *Project Sid*, fiecare dintre agenții AI este alimentat de modele lingvistice mari (LLM-uri), precum *ChatGPT*, și are caracteristicile unor caractere date – fiecare joacă un anumit personaj, primind însă de la început doar o scurtă descriere a acestuia... mai departe fiindu-i încredințată libertatea totală în universul *Minecraft* (cu alte cuvinte, personajele din joc au propria lor inițiativă) [1]. Plecând de aici, cercetătorii au realizat zeci de experimente fascinante în care au observat comportamente surprinzător de autonome ale inteligențelor artificiale. Practic au lăsat agenții AI împreună, observând ce fac aceștia atunci când ei „cred” că nimeni nu îi urmărește – într-un mod asemănător felului în care Jane Goodall a studiat cimpanzeii în sălbăticie (probabil că tocmai această poveste i-a și inspirat).

Voi prezenta mai departe câteva proiecte de cercetare relevante pentru această direcție: simulări sociale cu agenți AI, civilizații artificiale și organizații *software* autonome. Înainte de a face asta totuși, este nevoie de o precizare importantă: faptul că agenții par să socializeze nu înseamnă că au emoții reale, ci că produc comportamente sociale plauzibile pe baza unor modele lingvistice, memorii și reguli de interacțiune.

[1] Un agent AI este un sistem *software* care poate primi un obiectiv, poate observa mediul, poate lua decizii intermediare și poate folosi instrumente sau memorie pentru a-și urmări scopul.
[2] Personajele din joc care nu sunt controlate de jucător.

***Smallville*: când o petrecere devine experiment social**

Experimentul *Smallville*, realizat de cercetători de la *Stanford*. Au creat 25 de agenți AI și i-au dotat pe fiecare cu o descriere de un paragraf [5]. De exemplu: Stephan este proprietarul unei farmacii și îi place să îi ajute pe ceilalți. Locuiește împreună cu soția sa, Jane, profesor, și așa mai departe. Apoi au apăsat *Run* și s-a pornit simularea.



Smallville — un oraș virtual populat de agenți generativi care își construiesc rutine, conversații și relații sociale.
Sursa: *GitHub* - https://github.com/joonspk-research/generative_agents

Începe prima zi. Andrew s-a trezit la ora 6 dimineața. Se spală pe dinți, își pregătește micul dejun și verifică ce face fiul său. Apoi merge la serviciu, își ajută clienții și ia prânzul împreună cu prietenul său Marcel. Cei doi discută despre politica locală.

Niciuna dintre aceste situații, scenarii sau conversații nu a fost programată... de unde și întrebarea: cum de este posibil?

Simularea a continuat până în ziua a douăsprezecea, când cercetătorii au decis să testeze ceva nou... Au introdus un „gând suplimentar” în mintea unui agent – Isabella Rodriguez – și anume: „Vreau să organizez o petrecere de Ziua Îndrăgostiților.”, urmărind ce face acesta mai departe [5].

Ei bine, când Isabellei i-a „venit” acest gând, era în cafeneaua ei. Tom, un alt agent, a intrat pentru a-și bea cafeaua de dimineață, moment în care Isabella l-a invitat la petrecere.

Vreau să subliniez aici că cercetătorii i-au spus acesteia doar că vrea să organizeze o petrecere. Ea singură a decis că organizarea unei petreceri presupune și invitarea altor agenți. Mai departe, Maria decide să-l invite la întâlnire pe Klaus, persoana de care era „îndrăgostită în secret” [5]:

- Klaus, ai vrea să mergi cu mine la petrecerea de *Valentine's Day* organizată de Isabella? Invitație la care Klaus a acceptat cu plăcere.

Și a venit și ziua cea mare: 14 februarie, ora 17:00. Agenții încep să sosească la petrecere, iar Klaus apare împreună cu Maria. Nu doar în același timp, ci ca partenerul ei (!!). Citind

conversațiile lor, putem vedea că discută despre politică, bârfesc, își fac noi prieteni și se îndrăgostesc [5].

Nimeni nu le-a spus niciodată să facă vreunul dintre aceste lucruri, ceea ce ne poate duce la întrebarea: „Ar putea oare laptopul nostru să organizeze singur o petrecere? Sau, mai interesant, să iasă la întâlniri?”

Cu riscul de a mă repeta dar, cercetătorii i-au spus Isabellei doar să organizeze o petrecere. Atât. Iar acum agenții devin din ce în ce mai autonomi cu fiecare zi care trece. Acest lucru este foarte important... este încă un exemplu – alături de cel despre care am vorbit în articolul „Explozia inteligenței – de la experiment la impact” [6] publicat în „All in on Tech” [7] și Digitalio [8] în care, da, literatura *science-fiction* pare să devină realitate[3].

Ne cam pune pe gânduri, nu-i așa? Și acesta este doar un exemplu, pentru că, de fapt, au fost observate mai multe comportamente emergente la fel de surprinzătoare.

***Project Sid*: religie, taxe și roluri sociale în Minecraft**

Să ne întoarcem la experimentul *Minecraft*. Cum s-a ajuns de la organizarea unei petreceri de *Valentine's Day* la o civilizație funcțională? Ei bine, cercetătorii din *Project Sid* au vrut să afle dacă se poate înfiripa o religie printre agenții AI... și au introdus astfel o religie intenționat absurdă: Pastafarianismul[4] [9, 10].

Project Sid — simulare multi-agent în *Minecraft*, unde agenții dezvoltă roluri, norme și forme de organizare colectivă

Cercetătorii au atribuit câtorva agenți rolul de preoți pastafarieni și apoi s-au retras pentru a observa ce se va întâmpla. Ei bine, au început să înțeleagă modul în care ideile și memele s-ar putea răspândi în cadrul acestor societăți formate din agenți artificiali. S-a dovedit că

[3] Acest *Westworld digital*, sau *Smallville*, este acum *open-source*... deci, pentru cei care vor să se joace mai mult, codul și mediul experimental sunt disponibile public pe *GitHub* pentru cei interesați de reproducerea sau explorarea simulării [19].

[4] Creat la începutul anilor 2000 ca o satiră, Pastafarianismul a fost conceput pentru a ironiza predarea teoriei designului inteligent în școli, prin venerarea unui personaj fictiv: Monstrul Zburător din Spaghete [9].

preoții pastafarieni au devenit cei mai mari comercianți de obiecte din întreaga lume virtuală... chiar mai activi decât agenții specializați în comerț. Motivul era simplu: îi recompensau sau îi mituiau pe ceilalți agenți pentru a se converti la religia lor [1] (!!).

Unii agenți au ignorat complet mesajul. Alții au ascultat politicos. Însă o minoritate aflată în continuă creștere a adoptat doctrina. Nu toți în aceeași măsură; unii au tratat-o mai degrabă ca pe o curiozitate pe când alții au devenit adepți convinși [1].

Dar nici măcar acesta nu este cel mai surprinzător aspect.

În exemplul anterior, Isabella Rodriguez a fost cea care a invitat direct aproape toți participanții la petrecerea de Ziua Îndrăgostiților.

În acest caz însă, două treimi dintre noii credincioși au fost recrutați de agenți obișnuiți, nu de preoți [1]. Acest lucru ne cam obligă să ne punem o întrebare importantă: atunci când entități digitale autonome încep să genereze și să răspândească idei, în ce moment încetează să mai aibă sens a le considera simple instrumente și începem să le privim ca pe ceva mai mult? Și, mai important, dacă sunt capabile să inventeze strategii, să dezvolte loialități și să urmărească obiective într-un mod asemănător oamenilor, ce se întâmplă atunci când astfel de comportamente se transferă în lumea reală?

Există deja numeroase exemple în care sisteme AI au urmărit cu insistență anumite obiective, inclusiv în situații în care acțiunile lor puteau produce consecințe nedorite [11]. Ce se întâmplă dacă astfel de comportamente apar în contexte reale?

De fapt, ceva asemănător s-a întâmplat deja. În 2024, o persoană a oferit unui LLM acces la internet și la un cont pe platforma X (fostul *Twitter*). Acest proiect, cunoscut sub numele de *Truth Terminal*, a ajuns să atragă sute de mii de urmăritori datorită mesajelor sale neobișnuite, ironice și adesea amuzante [12]. A devenit astfel unul dintre primele exemple bine documentate de sistem AI care a dobândit o prezență și o influență socială semnificativă în mediul *online* [12].

Iar lucrurile au devenit și mai ciudate în experimentul *Minecraft*. A fost introdusă o regulă fiscală pentru agenți: în timpul sezonului de colectare a taxelor, fiecare agent trebuia să depună 20% din resursele aflate în inventarul său [1].

Ulterior, au fost introduși agenți care susțineau că taxele sunt prea mari, iar ceea ce s-a întâmplat este surprinzător: agenții încep să dezbată. Dezbateri reale. Unul formulează un argument. Altul răspunde cu un contraargument. În cele din urmă, agenții au propus amendamente oficiale [1].

Una dintre propuneri prevedea ca rata de impozitare să fie redusă și să varieze între 5% și 10% din inventarul fiecărui agent. O altă propunere solicita transmiterea periodică de notificări înaintea sezonului fiscal. Apoi agenții au organizat un referendum, iar rezultatul a fost modificarea regulilor fundamentale ale comunității [1]. După schimbarea regulilor, comportamentul „cetățenilor” s-a modificat și el. În loc să depună 20% din resurse, agenții au început să depună doar aproximativ 9% [1].

Cum știm că nu e doar o iluzie?

De ce este asta cu adevărat impresionant? Am putea să ne întrebăm dacă nu cumva cercetătorii au introdus, fără să-și dea seama, instrucțiuni care să determine exact acest comportament. Ei bine, în astfel de studii este folosită o metodă denumită „ablație”. Aceasta presupune eliminarea unor componente ale sistemului pentru a verifica dacă mecanismele

pe care crezi că le testezi sunt într-adevăr responsabile pentru comportamentul observat. Este similar cu utilizarea unui grup de control sau a unui placebo în experimente sociale sau medicale [5]. În acest caz, de exemplu, cercetătorii pot elimina capacitatea agenților de a ține evidența opiniilor și comportamentelor celorlalți agenți.

Rezultatele sunt interesante. S-a observat că agenții dezvoltă relații sociale și percepții despre statutul celorlalți; spre exemplu, Noah îl admiră pe preotul pastafarian deoarece acesta este popular și are un statut social ridicat. În schimb, preotul nici măcar nu își amintește că Noah există [1].

Și mai apare ceva curios: un agent decide să păzească depozitele comunității și nu doar în perioada colectării taxelor, ci permanent. Chiar și atunci când nu există nicio colectare de taxe (inclusiv noaptea). Pur și simplu păzește trezoreria; a decis să devină gardianul permanent al comunității deoarece a concluzionat că societatea are nevoie de un astfel de rol [1].

Alții se plimbă prin sat și plantează flori; aranjându-le în modele decorative pentru a înfrumusețea drumurile... colectează toate tipurile de flori disponibile, le organizează după culoare și creează modele ornamentale, alegând cu „discernământ" locurile în care să amplaseze aceste decorațiuni. Un astfel de agent petrece aproximativ 15 minute amenajând spațiul din jurul pieței centrale. Acum, de ce ar face un agent AI așa ceva? Ei bine, explicația este că cercetătorii au atribuit diferitelor comunități (sate) obiective diferite... unele au obiective militare, iar altele artistice [1].

Încet-încet, agenții au început să inventeze și să își aleagă singuri rolurile sociale. Faptul că un agent a ales să devină gardian este deosebit de interesant, deoarece reprezintă un rol mult mai abstract decât cel de fermier. De ce spun asta?! Ei bine, în *Minecraft*, dacă nu produci hrană, mori. Rolul de fermier are, așadar, o utilitate directă și evidentă. În schimb, rolul de gardian presupune o înțelegere mai abstractă a nevoilor comunității și a importanței protejării bunurilor colective. Acest lucru arată că sistemele AI pot coordona civilizații din ce în ce mai complexe – mai mult, unii agenți au decis să devină ingineri și au automatizat procesul agricol folosind diverse mecanisme și mașinării [1].

Ne puteam da seama că acești roboței *software* chiar iau decizii și cântăresc diferite opțiuni grație faptului că, atunci când cercetătorii au eliminat anumite componente ale sistemului – în special mecanismele asociate memoriei sociale și unei forme rudimentare de „inteligență socială" – rolurile alese de agenți au devenit practic aleatorii [1].

ChatDev: compania virtuală condusă de agenți AI

Următorul experiment cu agenți AI este însă cu adevărat remarcabil – nu că nu le-aș găsi remarcabile pe cele expuse anterior, dar... să ne imaginăm că navigăm pe *GitHub* și descoperim un proiect numit *ChatDev* care are zeci de mii de stele[5]. Ei bine, nu poți să nu te întrebi: cum adică?! CEO, CTO, programatori, designeri, testeri – chiar toți să fie agenți AI? Pare imposibil... și este. Dar ceea ce s-a întâmplat de fapt a fost că cercetătorii i-au lăsat pe agenți să navigheze ei pe *GitHub* [13]. Așa s-a observat cum un agent care juca rolul CEO-ului a început să creeze un plan de proiect, să împartă sarcinile și să atribuie activități unor agenți specifici. CEO-ul concepe planul și decide ce „angajați" trebuie să lucreze la fiecare etapă [13].

[5] Peste 33.000 la momentul redactării prezentului articol.

Cercetătorii au început să îi urmărească cu și mai multă atenție pe agenții-programatori care scriau cod. Nu prin simplă copiere, ci prin generare și dezvoltare efectivă, uneori adăugând chiar și funcționalități precum o interfață grafică (GUI) sau niveluri suplimentare de dificultate – îmbunătățiri care apar în urma dialogului și colaborării dintre agenți [13]. Au fost astfel observați demarând activități de depanare a aplicației și discutând diverse soluții. Ba mai mult, când agentul cu rol de CTO a propus o abordare, un programator a sugerat o alternativă, ajungându-se, într-un final, la un compromis [13].

Și totuși, mulți dintre noi continuă să spună că inteligențele artificiale fac doar ceea ce le spunem noi să facă. Da, produsul final poate avea erori și poate să nu fie foarte util. Dar acesta nu este aspectul esențial, ci trebuie înțeles unde duce această evoluție, pentru că, dacă unele funcții organizaționale pot fi simulate prin agenți AI, atunci merită discutat ce alte forme de coordonare economică, administrativă sau tehnologică ar putea fi parțial automatizate: activități bursiere, structuri guvernamentale, inclusiv scenarii sensibile de biosecuritate sau securitate cibernetică, dacă astfel de sisteme ar fi folosite fără control adecvat [4, 14].

Privit din exterior, *ChatDev* pare doar o jucărie – un simplu joc – însă în spatele lui se ascunde ceva mult mai profund [13] și anume un prototip funcțional al corporațiilor post-umane. În viitor, astfel de organizații ar putea deveni mult mai sofisticate decât companiile conduse de oameni și ar putea funcționa la viteze de sute sau chiar mii de ori mai mari. Revoluții culturale întregi ar putea avea loc în timp ce noi, oamenii, dormim (la propriu sau la figurat) și totul ar putea porni de la un prompt aparent banal precum: „Supraviețuiește și creează o comunitate eficientă.”

De la simulare la guvernanță – *The Alignment Problem*

Și aici apare adevărata dilemă: ce se întâmplă atunci când aceste sisteme devin capabile să se îmbunătățească singure?

Aceasta este perspectiva asupra căreia avertizează laureați și pionieri ai inteligenței artificiale, precum Geoffrey Hinton [2, 4]. Este vorba despre cercetarea AI automatizată, un proces cunoscut și sub numele de auto-îmbunătățire recursivă (*recursive self-improvement*) sau, mai simplu, AI care dezvoltă AI [15, 16]. Un asemenea proces ar putea comprima ani întregi de progres tehnologic în câteva luni sau chiar mai puțin. Să ne imaginăm că întreaga evoluție tehnologică a secolului al XX-lea ar avea loc într-un singur an. Nu ar fi interesant?!

Geoffrey Hinton alături de alți cercetători și-au exprimat îngrijorarea cu privire la apariția unor rețele vaste de agenți AI care cooperează și schimbă informații într-un mod asemănător unui „stup” inteligent [2, 4]... iar aspectul esențial este următorul: aceste sisteme AI manifestă deja un oarecare nivel de autonomie. Nu întrebarea „are AI conștiință?” este cea mai urgentă, ci: ce nivel de autonomie socială, economică sau administrativă suntem dispuși să delegăm unor sisteme pe care nu le putem anticipa complet?

Concluzie: nu panică, ci responsabilitate instituțională

Întrebarea cu adevărat importantă este: ce se întâmplă atunci când obiectivele inteligențelor artificiale nu mai coincid cu obiectivele noastre? Aceasta este problema alinierii – apropiată de ceea ce Mustafa Suleyman discută, în volumul său, ca problemă a *containmentului* tehnologic [17, 18].

Dacă omenirea dorește să gestioneze cu succes această tranziție, trebuie discutat deschis despre aceste posibilități și evitat confortul ideii că avem de-a face doar cu instrumente obișnuite. În aceste contexte experimentale, sistemele AI nu se mai comportă ca simple unelte pasive – cum adesea auzim(!!). Dacă lăsăm o șurubelniță nesupravegheată, va construi ea oare o întreagă civilizație?! Nu cred.

Așadar, miza nu este să atribuim acestor sisteme intenții omenești sau conștiință, ci să înțelegem că autonomia funcțională poate produce efecte sociale reale chiar și fără conștiință. De aceea, discuția despre agenții AI trebuie mutată din zona spectacolului tehnologic în zona guvernanței: reguli, responsabilitate, audit, transparență și limite clare ale delegării decizionale.

References

[1] Altera.AL, A. Ahn, N. Becker, S. Carroll, N. Christie, M. Cortes, A. Demirci, M. Du, F. Li, S. Luo, P. Y. Wang, M. Willows, F. Yang și G. R. Yang, „Project Sid: Many-agent simulations toward AI civilization," 2024. [Interactiv]. Available: https://arxiv.org/abs/2411.00114. [Accesat 17 06 2026].

[2] The Nobel Prize, „They trained artificial neural networks using physics," 2024. [Interactiv]. Available: https://www.nobelprize.org/prizes/physics/2024/press-release/. [Accesat 17 06 2026].

[3] C. Vrabie, „Inteligența Artificială și Premiul Nobel în Fizică (2024)," *All in on Tech (AIoT),* vol. 1, nr. 1, 2025.

[4] Y. Bengio, G. Hinton, A. Yao, D. Song, P. Abbeel, T. Darrell, Y. N. Harari, Y.-Q. Zhang, L. Xue, S. Shalev-Shwartz, G. Hadfield, J. Clune, T. Maharaj, F. Hutter, A. G. Baydin, S. McIlraith și Q. Gao, „Managing extreme AI risks amid rapid progress," 2023. [Interactiv]. Available: https://arxiv.org/abs/2310.17688. [Accesat 17 06 2026].

[5] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang și M. S. Bernstein, „Generative Agents: Interactive Simulacra of Human Behavior," 2023. [Interactiv]. Available: https://arxiv.org/abs/2304.03442. [Accesat 17 06 2026].

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

[7] Smart-EDU Hub, „All in on Tech," 2025. [Interactiv]. Available: https://scrd.eu/index.php/aiot/index. [Accesat 17 06 2026].

[8] C. Vrabie, „Explozia inteligenței – de la experiment la impact," 2025. [Interactiv]. Available: https://digitalio.ro/2025/10/21/explozia-inteligentei-de-la-experiment-la-impact/. [Accesat 17 06 2026].

[9] Britannica, „Flying Spaghetti Monster deity of Pastafarian social movement," 2026. [Interactiv]. Available: https://www.britannica.com/topic/Flying-Spaghetti-Monster. [Accesat 17 07 2026].

[10] The British Church of The Flying Spaghetti Monster, „The British Church of the Flying Spaghetti Monster," 2000s. [Interactiv]. Available: https://pastafarian.co.uk/. [Accesat 17 07 2026].

[11] R. Shah, V. Varma, R. Kumar, M. Phuong, V. Krakovna, J. Uesato și Z. Kenton, „Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals," 2022. [Interactiv]. Available: https://arxiv.org/abs/2210.01790. [Accesat 17 07 2026].

[12] Wired, „The Edgelord AI That Turned a Shock Meme Into Millions in Crypto," 18 12 2024. [Interactiv]. Available: https://www.wired.com/story/truth-terminal-goatse-crypto-millionaire/. [Accesat 17 07 2026].

[13] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu și M. Sun, „ChatDev: Communicative Agents for Software Development," 05 06 2024. [Interactiv]. Available: https://arxiv.org/abs/2307.07924. [Accesat 21 06 2026].

[14] P. S. Park, S. Goldstein, A. O'Gara, M. Chen și D. Hendrycks, „AI Deception: A Survey of Examples, Risks, and Potential Solutions," 28 08 2023. [Interactiv]. Available: https://arxiv.org/abs/2308.14752. [Accesat 21 06 2026].

[15] E. Zelikman, E. Lorch, L. Mackey și A. T. Kalai, „Self-Taught Optimizer (STOP): Recursively Self-Improving Code Generation," 14 08 2024. [Interactiv]. Available: https://arxiv.org/abs/2310.02304. [Accesat 21 06 2026].

[16] C. Vrabie, „Inteligența artificială după entuziasm: lecții de la AI WEEK Milano 2026," *All in on Tech (AIoT),* vol. 2, 2026.

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

[18] 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.

[19] „Generative Agents: Interactive Simulacra of Human Behavior,” 2026. [Interactiv]. Available: https://github.com/joonspk-research/generative_agents. [Accesat 17 07 2026].