RAPORT ȘTIINȚIFIC ETAPA 1 / 2025

Proiect: ***Norme de aserțiune în interacțiunea lingvistică dintre oameni și sisteme de inteligență artificială*** **(en: Norms of Assertion in linguistic Human-AI Interaction)**

Acronim: NIHAI

Cod proiect: ERANET-CHANCE-CRISIS-NIHAI

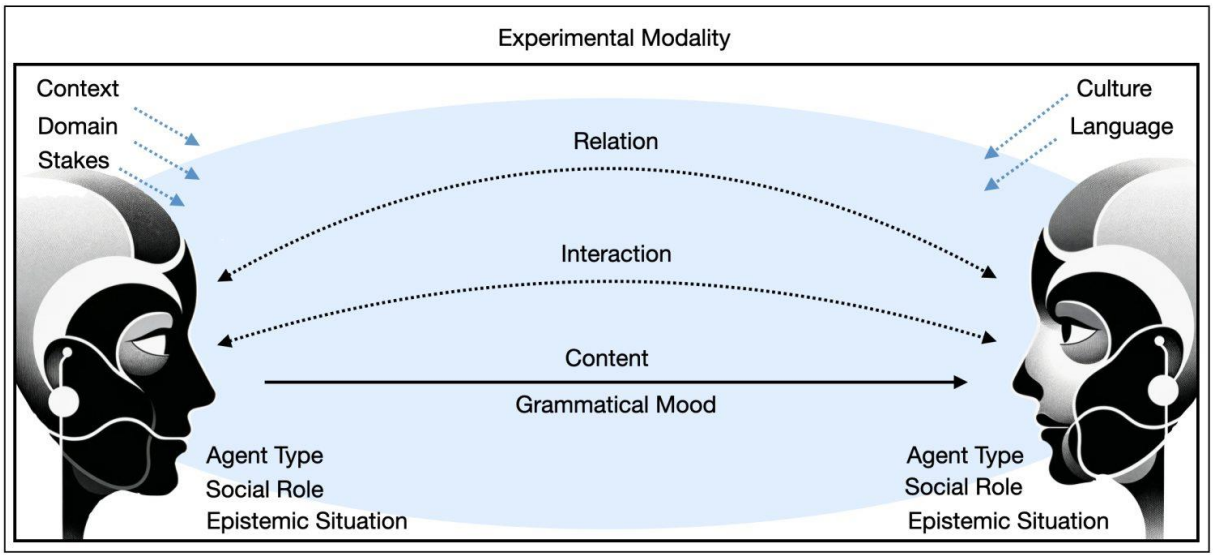
Contract: nr. 83 din 01/03/2025

Perioada de implementare: 2025-2027

Etapa raportată: Etapa 1 – 2025 (*Fundamentare conceptuală – sistematizarea literaturii de specialitate privind normele de comunicare lingvistică în interacțiunea dintre oameni și IA*)

Director proiect: Mihaela Constantinescu

Instituție coordonatoare: Universitatea din București, Facultatea de Filosofie, Centrul de Cercetare în Etică Aplicată (CCEA)



Rezumat executiv

În anul 2025, partenerul român din consorțiul proiectului *NIHAI – Norme de aserțiune în interacțiunea lingvistică dintre oameni și sisteme de inteligență artificială* a avut ca principal obiectiv de cercetare fundamentarea conceptuală și sistematizarea literaturii de specialitate privind normele de comunicare lingvistică în interacțiunile om–inteligență artificială.

O primă direcție de cercetare în acest sens a analizat interacțiunile tinerilor cu companioni artificiali bazați pe modele lingvistice mari (boți conversaționali sociali). Deși aceste interacțiuni pot oferi beneficii emoționale, ele pot amplifica vulnerabilități precum dependența afectivă, confuzia ficțiune-realitate sau așteptări nerealiste. Cercetarea a propus un cadru de responsabilitate care impune companiilor să asigure: claritatea limitelor tehnologiei, evitarea iluziei reciprocității emoționale, semnalizarea naturii ficționale a interacțiunii și protejarea utilizatorilor vulnerabili.

O a doua direcție de cercetare a explorat normele relaționale din cooperarea om-IA, analizând cum reguli precum încrederea, loialitatea sau reciprocitatea funcționează diferit când partenerul este un sistem artificial fără experiențe subiective. Cercetarea propune un cadru de „norme relaționale" adaptate cooperării om-IA, care clarifică responsabilitățile agentului artificial, distribuirea răspunderii morale între utilizatori, dezvoltatori și instituții, și ajustarea așteptărilor de comunicare.

Activitățile din această etapă au inclus participarea la două întâlnirii de consorțiu, elaborarea și revizuirea articolelor științifice, dezvoltarea platformei publice de comunicare a proiectului și participarea la conferințe și evenimente internaționale. Rezultatele majore constau în trimiterea spre publicare în reviste internaționale a două articole de cercetare (*The Sorrows of Young Chatbot Users: Harm and Responsibility in Human–AI Relationships* și *Relational Norms for Human–AI Cooperation*), redactarea a două articole de cercetare în vederea publicării în reviste internaționale (*Functional Fictionalism* și *Dark Moves in Human–AI Interaction*), precum și stabilirea cadrului metodologic pentru studiile experimentale planificate în 2026.

Echipa a contribuit, de asemenea, la diseminarea publică a cercetării prin conferințe academice și evenimente de outreach în Marea Britanie, Spania și România, consolidând poziția consorțiului în dezbaterile internaționale privind etica comunicării om–IA.

Gradul general de realizare al etapei este complet, iar activitățile derulate asigură o bază solidă pentru faza experimentală din 2026.

1. Descrierea generală a etapei

Etapa I a proiectului *NIHAI – Norme de aserțiune în interacțiunea lingvistică dintre oameni și sisteme de inteligență artificială* s-a derulat în perioada 1 martie–31 decembrie 2025 și a avut ca scop principal **fundamentarea conceptuală și sistematizarea literaturii de specialitate privind normele de comunicare lingvistică în interacțiunile om–inteligență artificială**.

Obiectivele de cercetare ale etapei:

- Studiul conceptelor cheie - normele de comunicare lingvistică în interacțiunea dintre oameni și inteligența artificială (IA)
- Sistematizarea literaturii de specialitate - interacțiunea om-IA

Plan de realizare/ etapa 1	**Livrabile realizate/ etapa 1**
site proiect și canale social media	crearea și administrarea site-ului proiectului și a canalelor social-media (LinkedIn, Bluesky, X/Twitter) - https://talkingtobots.net/ și https://www.ccea.ro/nihai/
1 intalnire de consorțiu (Graz)	2 întâlniri de consorțiu pentru coordonarea partenerilor și stabilirea planului comun de cercetare (06/2025, Graz, Austria; 11/2025, Nottingham, UK)
1 articol de sistematizare a literaturii de specialitate privind normele de comunicare lingivistică în interacțiunea dintre oameni și IA	2 articole de cercetare în evaluare / acceptate pentru publicare; 2 articole cercetare în pregătire (*preprint*);
registru online public de date asociat proiectului	OSF repository cu înregistrarea articolelor experimentale din proiect
participare la un eveniment non-academice de promovare a rezultatelor cercetării	2 evenimente de cercetare 4 evenimente publice non-academice

Gradul de implementare al etapei este **complet**, toate activitățile planificate fiind realizate în termenii prevăzuți în planul de implementare, iar rezultatele obținute contribuind direct la atingerea obiectivelor generale ale proiectului.

2. Rezumatul activităților și al progresului științific

Echipa română a consorțiului NIHAI a contribuit la conturarea cadrului conceptual al proiectului prin cercetare teoretică, redactarea și revizuirea articolelor științifice, participarea la conferințe internaționale și activități de diseminare publică.

2.1. Articole științifice (publicate / în evaluare)

1. **„The Sorrows of Young Chatbot Users: Harm and Responsibility in Human–AI Relationships”** (în curs de publicare, revista internațională *Topoi*) – Cristina Voinea, Christopher Register, Sebastian Porsdam Mann, Julian Savulescu și Brian D. Earp (Voinea coautoare în echipa NIHAI)

Lucrarea examinează interacțiunile cu compani-oni artificiali bazați pe LLM-uri ca forme de angajament ficțional, comparându-le cu lectura sau jocurile video și analizând responsabilitatea morală a companiilor dezvoltatoare pentru eventualele prejudicii produse utilizatorilor. Studiul argumentează că, deși aceste interacțiuni pot avea beneficii emoționale sau terapeutice, ele pot totodată amplifica vulnerabilități — cum ar fi dependența afectivă, confuzia dintre ficțiune și realitate sau formarea unor așteptări nerealiste cu privire la capacitățile și intențiile sistemelor. În acest context, este propus un cadru de responsabilitate care pune accent pe modul în care companiile trebuie să proiecteze și să reglementeze comunicarea dintre oameni și agenți artificiali: claritatea asupra limitelor tehnologiei, evitarea iluziei reciprocității emoționale, semnalizarea corectă a naturii ficționale a interacțiunii și implementarea unor mecanisme menite să protejeze utilizatorii vulnerabili. Lucrarea susține că normele de comunicare om–IA nu reprezintă doar o chestiune tehnică sau stilistică, ci una morală fundamentală, întrucât modul în care IA „vorbește” cu oamenii poate influența direct bunăstarea și autonomia acestora.

2. **„Relational Norms for Human–AI Cooperation”**, Earp et al. (Mihaela Constantinescu și Cristina Voinea coautoare din echipa NIHAI) – *în evaluare*

Articolul abordează normele relaționale care guvernează cooperarea omului cu IA. Sunt discutate situații în care reguli precum încrederea, loialitatea, responsabilitatea partajată sau așteptările de „reciprocitate” funcționează diferit atunci când partenerul de cooperare este un sistem artificial, incapabil de experiențe subiective. Pe această bază, articolul propune un cadru de „norme relaționale” adaptate cooperării om–IA, care să clarifice ce fel de responsabilități poate avea un agent artificial, cum ar trebui distribuită răspunderea morală între utilizatori, dezvoltatori și instituții și în ce fel trebuie ajustate așteptările noastre de comunicare și cooperare. În acest sens, contribuția subliniază că dezvoltarea IA cooperativă nu poate ignora întrebarea: ce fel de norme de comunicare și relaționare sunt adecvate pentru o „relație” cu entități artificiale menită să preia roluri umane?

2.2. Articole științifice (*preprint*)

1. **„Functional Fictionalism and the Nature of LLM Assertions”**, Zorilă, Voinea, Figura și Constantinescu – *preprint*

Lucrarea investighează modul în care utilizatorii ajung să trateze răspunsurile chatbot-urilor bazate pe modele mari de limbaj (LLM) ca afirmații epistemic semnificative, deși aceste sisteme nu dețin stări mentale, intenții sau capacitate de înțelegere. Pornind de la tensiunea dintre teoriile metasemantice dominante, care consideră limbajul produs de LLM-uri lipsit de sens literal și practicile curente de utilizare, articolul propune o teză dublă: (1) interacțiunea cu chatbot-urile se bazează pe un cadru ficționalist de tip „make-believe”, prin care utilizatorii tratează sistemele ca interlocutori; iar (2) comportamentul lingvistic al LLM-urilor simulează funcțional actul aserțiunii, generând aparența unei intenții de a spune adevărul și activând mecanisme de încredere specifice mărturiei umane.

2. **„Dark Moves in Human–AI Interaction”**, Tofan,Voinea, Dancu, Figura, Skokzen, Kandul, Soitu, Țânculescu-Popa, Ceocea, Zorilă, și Constantinescu – *preprint*

Analiză sistematică a literaturii privind strategiile conversaționale manipulative sau dăunătoare („dark moves”) utilizate de sistemele LLM. Studiul cartografiază tipuri recurente de astfel de strategii – de la presiuni subtile asupra deciziilor utilizatorilor și formulări care induc dependență sau conformism, până la prezentarea selectivă sau înșelătoare a informațiilor – și arată că, în pofida îngrijorărilor crescânde, cercetările existente folosesc adesea noțiuni vagi și metodologii eterogene. Lucrarea identifică lacunele conceptuale și metodologice din acest domeniu și argumentează pentru nevoia unui vocabular mai precis și a unor protocoale standardizate de evaluare a „dark moves” în interacțiunile om–IA. În acest cadru, normele de comunicare capătă un rol central: pentru a putea preveni și reglementa astfel de strategii, este necesar un set explicit de principii privind onestitatea, claritatea, evitarea presiunii psihologice și protejarea utilizatorilor vulnerabili în conversațiile cu sistemele AI.

3. Prezentarea rezultatelor cercetării

Echipa română NIHAI a participat activ la întâlnirile de consorțiu NIHAI, precum și la evenimente științifice și de popularizare, contribuind la vizibilitatea proiectului și la promovarea dezbaterilor etice privind comunicarea om–IA, precum și la crearea de punți între mediul academic și cel public privind utilizarea responsabilă a IA conversaționale.

3.1 Întâlniri de consorțiu

1. **Kick-off conference - „AI Assertion”,** Universitatea din Graz, Austria, 22-23.06.2025.

Conferință interdisciplinară de lansare a proiectului NIHAI, care examinează natura aserțiunii în IA prin perspectivele filosofiei, științelor cognitive, lingvisticii și interacțiunii om-IA. Conferința analizează când și cum pot fi considerate sistemele de IA ca asertori și cum se compară acestea cu actele de vorbire umane; dacă normele conversaționale existente se aplică IA sau dacă sunt necesare standarde noi; responsabilitate și asumarea răspunderii în comunicarea mediată de IA, în special în cazurile de neînțelegere sau înșelăciune; modul în cae funcționează concepte precum încrederea și normativitatea în interacțiunea om-IA și dacă încrederea în utilizarea limbajului de către IA este justificată sau greșit plasată.

2. **Conferința „Knowledge Exchange for Slow Hope”** organizată la Nottingham în perioada 24-25.11.2025

Eveniment care a reunit toate echipele de cercetare HERA/CHANSE Crisis pentru două zile de schimb de idei, reflecție și consolidare comunitară. Întâlnirea a oferit o oportunitate valoroasă de a împărtăși evoluțiile din cadrul proiectelor și de a întări spiritul colaborativ care animă această rețea extinsă de cercetare.

3.2. Conferințe și prezentări științifice

1. ***“Can We Talk? The Ethics of Human–AI Communication”***, *LLMs @ Oxford*, poster talk, Department of Computer Science, University of Oxford, 14 septembrie 2025.

Cristina Voinea discută în această prezentare normele și responsabilitățile care ar trebui să ghideze interacțiunea dintre oameni și sistemele de inteligență artificială. Cristina examinează felul în care comunicarea cu modele lingvistice avansate modelează percepțiile, comportamentele și deciziile utilizatorilor și argumentează în favoarea unor reguli etice ferme privind transparența, limitarea potențialului de manipulare, respectarea autonomiei umane și gestionarea realistă a așteptărilor. În esență, posterul pune în discuție cum putem construi un cadru de comunicare cu AI-ul care să fie nu doar eficient, ci și moral legitim, sigur și responsabil.

3. *"**How Should We Live Well with LLMs?"***, *AI for Flourishing Conference*, *keynote,* University of Navarra, 30 iunie 2025.

Mihaela Constantinescu oferă în această prezentare un răspuns la întrebarea *Cum ar trebui să conviețuim bine cu LLM-urile?* folosind cadrul eticii virtuții. În tradiția antică a eticii virtuții, modul în care cineva ar trebui să trăiască este fundamentat pe o concepție a înfloririi umane ce are în centrul său virtutea (*aretē*). Această perspectivă nu întreabă doar „ce este permis?" sau „ce maximizează utilitatea?", ci mai degrabă „ce fel de persoană ar trebui să fiu?" și „cum pot să înfloresc ca ființă umană?". Aplicată la interacțiunea cu LLM-urile, etica virtuții ne îndeamnă să evaluăm critic modul în care aceste tehnologii ne modelează caracterul și capacitățile cognitive. Astfel, utilizarea modelelor lingvistice mari pentru înflorirea umană trebuie să țină cont de viciile acestor modele – tendința lor de a genera informații plauzibile dar inexacte, de a reproduce prejudecăți, sau de a crea iluzia înțelegerii fără substanță reală. Mai important, trebuie să evităm delegarea unor sarcini cognitive esențiale – gândirea critică, creativitatea autentică, judecata morală, rezolvarea complexă de probleme – care ar putea duce, în cele din urmă, la o debilitare umană. Întrebarea centrală devine: folosim LLM-urile ca unelte care ne amplifică virtuțile – curiozitatea intelectuală, înțelepciunea practică, capacitatea de discernământ – sau le permitem să submineze aceste calități prin dependență excesivă și atrofiere cognitivă?

3.3. Activități de diseminare și comunicare publică

1. ***"Automated Moral Reasoning"*,** programul de popularizare *BiteSize Ethics 2025 – Ethics in the Age of AI*, Uehiro Oxford Institute, 13 august 2025.

Cristina Voinea pornește în această prezentare de la observația că oamenii tind să aibă încredere în AI pentru decizii practice, dar devin reticenți atunci când miza este morală, prezentarea subliniază că nu există încă un cadru matematic general acceptat pentru raționamentul etic, ceea ce face automatizarea moralității deosebit de problematică. Legătura cu normele de comunicare om–AI apare în accentul pus pe necesitatea ca sistemele să comunice clar limitele propriilor „capacități morale", incertitudinile și presupunerile pe care le folosesc, astfel încât utilizatorii să nu fie induși în eroare, să își poată păstra autonomia și să înțeleagă când își asumă ei înșiși responsabilitatea finală. În acest sens, prezentarea argumentează că orice proiect de automatizare a raționamentului moral trebuie însoțit de norme riguroase de transparență, onestitate și non-manipulare în interacțiunea cu utilizatorii.

2. Prezentare despre etica interacțiunilor om–IA în fața unei **delegații a *Department for Science, Innovation and Technology (DSIT)*, Guvernul Marii Britanii,** 30 septembrie 2025.

Cristina Voinea abordează în această prezentare provocările normative care însoțesc integrarea sistemelor inteligente în spațiul public și în procesele instituționale. Expunerea a

evidențiat riscurile legate de opacitatea modelelor avansate, de posibila influențare a utilizatorilor prin formulări persuasive și de tendința oamenilor de a supraestima competențele sistemelor automate. În acest context, a fost subliniată importanța stabilirii unor norme de comunicare clare între oameni și AI, norme care să impună transparență, delimitarea explicită a limitelor tehnologice, responsabilizarea utilizatorilor și evitarea oricăror forme de manipulare sau inducere în eroare. Prezentarea a insistat că dezvoltarea și implementarea IA la nivel guvernamental trebuie să fie însoțite de standarde robuste de comunicare etică, pentru a susține încrederea publică și decizii informate.

3. Discuție publică despre utilizarea responsabilă a LLM-urilor, evenimentul de business și inovație **IMPACT Hub Bucharest***,* 17 septembrie 2025

Mihaela Constantinescu discută alături de antreprenorul Dragoș Stanca despre expansiunea inteligenței artificiale generative în educație: cum amplificăm valorile umane fără a le eroda și cine poartă responsabilitatea pentru asta? Distincția dintre utilizare responsabilă și abuz devine esențială. Abuzul se manifestă prin fraudă academică, dar problema mai profundă este decăderea cognitivă. Studii recente, inclusiv de la MIT, arată că utilizarea excesivă a AI diminuează activitatea cerebrală și memoria. Dacă tinerii se bazează pe LLM-uri pentru a accesa cunoașterea fără efort intelectual propriu, riscăm atrofierea capacităților cognitive. Soluția constă în cultivarea gândirii critice – cea mai bună lentilă pentru utilizarea tehnologiei în secolul XXI. Cu discernământ critic, inteligența generativă devine un instrument excelent pentru înflorirea umană, nu pentru subjugare cognitivă. Educația trebuie să echilibreze inovația cu protejarea capacităților fundamentale, asigurând că tehnologia servește dezvoltarea umană autentică, nu o subminează.

4. **Masa rotundă „Cunoaștere în era inteligenței artificiale”**, un eveniment organizat în cadrul proiectului Biblioteca Alifanti și găzduit la Rezidența 9, programat pentru 5 decembrie 2025.

Alexandra Zorilă, bursier în cadrul proiectului, se alătură în cadrul aceste mese rotunde unui grup de cercetători, cadre universitare, artiști și profesioniști interesați de impactul tehnologiilor emergente asupra modului în care producția și circulația cunoașterii se transformă în contextul dezvoltării accelerate a modelelor mari de limbaj (LLM). Temele includ: statutul epistemic al outputurilor generate de LLM-uri, limitele și riscurile lor, modificările în dinamica expertizei, apariția noilor forme de „cunoaștere asistată” și implicațiile etice asociate utilizării unor sisteme non-agențiale în activități tradițional umane (analiză, sinteză, comunicare, creativitate).

5. Dificultăți și planuri pentru etapa următoare

Etapa I s-a desfășurat conform planificării generale, fără întârzieri majore. Principalele provocări au fost legate de coordonarea internațională a echipelor în procesul de redactare a articolelor și de timpii prelungiți ai procesului editorial (*peer review*).

Pentru etapa a II-a (2026), echipa română va:

- dezvolta designul studiilor experimentale online și de laborator;
- colabora la redactarea articolului empiric bazat pe aceste studii;
- continua publicarea rezultatelor teoretice începute în 2025;
- actualiza și extinde registrul public de date.

Director Proiect:
Mihaela Constantinescu
Universitatea din București – Facultatea de Filosofie