Raportul Științific 2025 pentru Contract nr. 66PHE/2024; etapă nr. 2/2025; Avataresponsibility: Influența avatarurilor bazate pe inteligență artificială asupra comportamentului moral.

A. **Descrierea științifică cu punerea în evidență a rezultatelor etapei anuale și gradul de realizare a obiectivelor**;

Începând cu anul 2024, Centrul de cercetare în etică aplicată (CCEA) al Universității din București găzduiește proiectul ERC Starting Grant avataResponsibility, dedicat explorării impactului utilizării avatarurilor (digitale, augmentate și robotice) asupra comportamentelor, practicilor și instituțiilor umane. Activitățile suport și direcțiile de dezvoltare propuse în acest proiect de Premiere Instituțională urmăresc poziționarea CCEA ca partener strategic în consorții europene pentru aplicații tip Horizon în domeniul eticii tehnologiilor emergente.

În perioada 2024-2027, proiectul de Premiere Instituțională își propune să atingă următoarele rezultate.

- 5 articole academice publicate în reviste internaționale prestigioase cu sistem de evaluare (peer-review) internationale (ISI, Philosopher's index, SCOPUS, MEDLINE, EBSCO);
- 10 mobilități internaționale ale membrilor echipei de proiect;
- 2 cursuri de specializare ale membrilor echipei de proiect;
- 3 evenimente internaționale organizate;
- dezvoltarea unui laborator cu tehnologii VR (realitate virtuală);
- acordarea a 2 burse de cercetare

Grad de implementare pentru anul 2025 (realizare obiective, atingerea rezultatelor preconizate): complet.

- 1 articol academic publicat în reviste internaționale prestigioase cu sistem de evaluare (peer-review) internationale (ISI, Philosopher's index, SCOPUS, MEDLINE, EBSCO);
- 3 participări la workshop-uri/conferințe internaționale;
- 1 intalnire informala cu lectori straini invitați
- 2 burse cercetare
- Dotare laborator cu tehnologii VR
- 1 eveniment stiintific cu participare interna si internationala
- 1 curs perfectionare

Sumar al progresului

B.1. Publicații:

1. **Mihailov, E. (2025). Enlightened Beneficence: A Kantian Alternative to Effective Altruism. Res Publica, 1-19. https://doi.org/10.1007/s11158-025-09727-w / Indexată în Arts & Humanities Citation Index (ISI), EBSCO, ERIH PLUS, Philosopher's Index.**

Rezumat: În această lucrare, dezvolt o alternativă Kantiană la altruismul eficcace, pe care o numesc binefacere iluminată. În primul rând, susțin că un cadru Kantian, bazat pe standardul universalizabilității și pe idealul umanității ca scop în sine, favorizează binefacerea imparțială. De fapt, Kant a anticipat ideea cercului extins al moralității ca o schimbare incluzivă către un punct de vedere care este imparțial la nivel universal. În al doilea rând, propun o distincție între eficacitatea absolută, aliniată la maximizarea utilitaristă, și eficacitatea relativă, care se concentrează pe optimizarea rezultatelor în cadrul unor constrângeri date și este integrată într-un cadru Kantian care combină pluralismul valorilor cu monismul procedural. Susțin că, deși evaluarea consecințelor este controversată la nivel fundamental în etica Kantiană, ea este inerentă îndeplinirii datoriei imperfecte a binefacerii. Considerațiile practice ale lui Kant despre ajutorul eficace reflectă o preocupare pentru impact și pentru luarea deciziilor bazate pe dovezi. În a treia secțiune, voi ilustra modul în care o abordare Kantiană lărgește sfera de aplicare a ajutorului global pentru a include intervenții care îmbunătățesc în mod eficient capacitatea de acțiune umană, nu doar bunăstarea oamenilor. În cele din urmă, susțin că sursele intelectuale pentru utilizarea rațiunii și a dovezilor în vederea îmbunătățirii societății se găsesc în mișcarea iluminismului și în ascensiunea științelor sociale experimentale.

Dobre, Mihnea. (2025). Clerselier's Programmatic Descartes. Early Science and Medicine 30 (2–3): 273–281. https://doi.org/10.1163/15733823-20251350. Indexată în Web of Science, Scopus, ERIH PLUS.

Rezumat: Perioada modernă timpurie este adesea descrisă ca o epocă în care corespondența a jucat un rol semnificativ în modelarea dezvoltării cunoașterii filozofice. Marile personalități din

Republica Literelor sunt prezentate în mod obișnuit ca exemple ale rețelelor complexe care au fost stabilite în întreaga Europă și care legau gânditori de tot felul. Unele dintre aceste personalități sunt lăudate pentru contribuțiile lor la înființarea unor instituții importante – de exemplu, Henry Oldenburg ca secretar al Societății Regale și editor al Philosophical Transactions –, în timp ce altele sunt celebrate pentru capacitatea lor de a traversa granițele disciplinare și confesionale prin schimbul de scrisori. Acest volum pregătit de Siegrid Agostini se referă la o astfel de personalitate: Claude Clerselier (1614–1684). Clerselier este de obicei înfățișat ca principalul promotor al cartezianismului în secolul al XVII-lea. El s-a ocupat de câteva ediții importante ale lui René Descartes. Articolul urmărește reconstrucția istorică a modului în care Claude Clerselier, editorul manuscriselor rămase nepublicate la moartea lui René Descartes, impune un anume tip de înțelegere pentru „filosofia carteziană". La un nivel mai general, problema analizată în această lucrare ține de felul în care reconstrucțiile istorice și categoriile istoriografice se împletesc în noile explicații cu privire la diferite concepte și doctrine din istoria filosofiei.

Constantin Vică, Radu Uszkai, Alexandru Dancu, Cristina Voinea, "Avatarobots and the Crisis of Democracy: From Bot Armies to Metaverse Swarms", sub evaluare la Internet Policy Review, indexată în Scopus, Emerging Sources Citation Index (ESCI), Directory of Open Access Journals (DOAJ).

Rezumat: În această lucrare susținem că mediile imersive ridică noi provocări pentru comunicarea democratică, transformând exprimarea politică în interacțiune întrupată. Analizăm modul în care avataroboții alimentați de AI pot fabrica majorități sintetice, simula agenție politică și distorsiona semnalele de care depinde legitimitatea democratică. Pentru a evalua aceste riscuri, dezvoltăm Principiul Exprimării Libere Responsabile (PAFE), structurat pe trei criterii: autenticitate, non-dominare și verificabilitate. Acestea corespund responsabilității pentru discurs, protecției împotriva dominației economice și originii verificabile a comunicării politice. Folosim PAFE pentru a analiza riscurile democratice generate de avataroboți și pentru a deriva ghiduri de proiectare și guvernanță pentru medii imersive care urmăresc să funcționeze ca sfere publice democratice.

Tofan, Cristina Maria. (2025). "Almighty AI! A brief review about beliefs about AI and instruments to (de)humanize AI", sub evaluare la Journal for the Study of Religions and Ideologies, idexată în ISI - Thompson-Reuters, Web of Knowledge, SCOPUS, EBSCO, ProQuest, Brill, Gale, ERIH Plus.

Rezumat: Articolul sintetizează contribuții teoretice și empirice din filozofie, psihologie socială, interacțiunea om–calculator și etica inteligenței artificiale, pentru a cartografia tipurile distincte de credințe despre IA, de la credințele bazate pe instrumente și pe mărturie, la cele bazate pe tehnologie, credințele populare și metaforele, credințele generate de conspirații, precum și credințele profesionale specifice unor domenii despre sfera de acțiune, autonomia și inteligența IA. Pe baza acestei hărți conceptuale, analizăm modul în care aceste credințe interacționează cu dispozițiile psihologice (de exemplu, evitarea incertitudinii, anxietatea, autoeficacitatea), imaginarul sociotehnic (de exemplu, IA ca actor epistemic) și contexte organizaționale, pentru a modela încrederea, percepția amenințării și disponibilitatea de a adopta IA. În paralel, trecem în revistă „instrumentele" care (de)umanizează IA, antropomorfizarea, atribuirea unei teorii a minții, încadrările privind responsabilitatea și autonomia, practicile de IA explicabilă și modelele formale de atribuire precaută a credințelor, care modulează dacă IA este tratată ca un simplu instrument, ca un agent cvasi-uman sau ca ceva intermediar. Folosind o sinteză narativă informată de lucrări teoretice recente în domeniul interacțiunii om–IA am organizat sinteza în jurul a trei întrebări:(1) Cum sunt structurate conceptual și măsurate credințele despre IA? (2) Prin ce mecanisme cognitive, afective și sociotehnice aceste credințe umanizează sau de-umanizează IA? (3) Cum se reflectă aceste credințe în reacțiile morale, epistemice și comportamentale față de IA, incluzând încrederea, vina și aplicarea normelor? Ne așteptăm ca această analiză să evidențieze legături sistematice între tipurile de credințe și practicile de (de)umanizare, să clarifice modul în care IA poate atât eroda, cât și susține autonomia epistemică umană și să propună un cadru conceptual pentru cercetări empirice viitoare asupra sistemelor de credințe despre IA în contexte cotidiene și profesionale.

B.2. Mobilități naționale și internaționale

Cristian Iftode, „Autonomia relaţională: un paradox?", prezentare susţinută în Conferințele București-Chișinău, Ediția a VI-a „Identitate şi valori culturale în context european" – 10-12.04.2025, Chişinău, Facultatea de Istorie și Filosofie a Universității de Stat din Moldova (https://filosofie.unibuc.ro/editia-a-vi-a-a-conferintelor-bucuresti-chisinau-cu-tema-identitate-si-valori-culturale-in-context-european-10-12-aprilie-2025/).

Rezumat: Prezentarea mea a discutat mai multe versiuni propuse în filosofia morală şi politică de dată recentă ale conceptului de autonomie relaţională. Miza a fost de a lămuri dacă aceste versiuni amendate ale ideii moderne de autonomie sunt mai adecvate pentru a da seama de realitatea socială contemporană şi importanţa relaţiilor personale în configurarea identităţii umane decât versiunile clasice ale ideii de autonomie, de tip substanţial sau procedural. Luând în discuţie cazuri-limită din registrul „opresiunii internalizate", al adicţiei asumate ca trăsătură identitară sau al vieţii de supunere liber aleasă, am încercat să înţeleg dacă noţiunea de autonomie relaţională poate da seama de aceste situaţii cel puţin la prima vedere paradoxale şi greu de compatibilizat cu o viziune de tip liberal. Sugestia mea a fost că pentru a depăşi aceste paradoxuri este nevoie să introducem şi o noţiune de autenticitate relaţională ireductibilă la conceptul de autonomie relaţională sau la o simplă condiţie necesară pentru exercitarea acesteia.

Daniel Nica, „Identitate și autenticitate. Trei modele ale autenticităţii", prezentare susţinută în cadrul Conferințele București-Chișinău, Ediția a VI-a Identitate şi valori culturale în context european – 10-12.04.2025, Chişinău, Facultatea de Istorie și Filosofie a Universității de Stat din Moldova (https://ibn.idsi.md/ro/collection_view/3561).

Rezumat:

În filosofie, există două modalități de problematizare a identității: pe de o parte, ca problemă metafizică, sub forma conceptului de identitate personală (concept intim legat de ideea identităţii numerice); pe de altă parte, ca problemă a filosofiei practice, sub forma conceptului de autenticitate. În literatura de specialitate dedicată acestei probleme a filosofiei practice există o dublă există o distincție standard între două modele ale autenticităţii: una „esenţialistă" (autenticitatea ca descoperire de sine, de inspiraţie rousseauistă) și alta „existenţialistă" (autenticitatea ca „alegere originară" a unui proiect fundamental, de inspiraţie kierkegaardiano-

sartriană). Primul, un model expresivist, asumă preexistenţa unui sine substanţial, „natural”, care trebuie descoperit în adâncurile unei interiorităţi privilegiate și actualizat prin exprimarea celor mai adânci dorinţe; al doilea, un model decizionist, pleacă de la premisa că un astfel de sine, deși nu există, poate fi constituit prin alegerile libere ale individului. Însă, în ciuda diferenţelor evidente, ceea ce au în comun cele două modele este ceea ce R. Rorty numea „căutarea unei purităţi a sinelui”. Ceea ce îmi propun să arăt este că putem vorbi și de un al treilea model, unul experimentalist, a cărui origine se găsește în filosofia lui Nietzsche, un model care are unele „asemănări de familie” cu celelalte două, dar nu este reductibil la acestea. Teza care sintetizează acest model este că „sinele autentic” se constituie într-o manieră estetică, experimentând succesiv o pluralitate de formule și practice identitare. Două idei secundare se desprind și acompaniază această teză: a. individul autentic reușește să armonizeze cât mai multe experienţe, valori și pasiuni într-un proiect de viaţă, coerent și original; b. individul autentic își integrează toate contradicţiile existenţiale într-o unitate dinamică, deschisă către alte reconfigurări ulterioare.

Dancu, A., Tofan, C. M. and Mihailov, E. (2025), WEIRD by design: How western moral values shape LLM ethical reasoning, prezentare susținută în cadrul International Conference on Themes and Fundamental Ideas in the Field of Psychology and Educational Sciences; Gheorghe Zane Institute for Economic and Social Research, Romanian Academy – Iasi Branch, Iași (Romania), 24 octombrie 2025.

Rezumat: Studiul analizează modul în care modelele mari de limbaj (LLM) încorporează și exprimă valori morale. Cercetarea evaluează 15 LLM-uri cu instrumente precum MFQ, MAC și SVS, fiecare testat de 100 de ori. Analizele statistice (ICC, chi-square, Cramér’s V) au vizat consistența și diferențele între modele, iar rezultatele sugerează o aliniere mai puternică la valorile individualiste occidentale și mai puțin la cele tradiționale sau sociocentrice. Studiul oferă perspective asupra bias-ului moral și al alinierii culturale în agenții artificiali, cu implicații pentru designul etic al AI.

Tofan, C. M. (2025), ”Trust, Betrayal, and Punishment: How Anthropomorphism Influences AI Consequences. An Experimental Research Protocol”, prezentare susținută în

cadrul Bucharest – Oxford Workshop in Practical Ethics; Faculty of Philosophy, Bucharest University, București (România) 10 aprilie 2025.

Rezumat: Pe măsură ce inteligența artificială (IA) devine tot mai integrată în procesul decizional uman, implicațiile etice și psihologice ale încălcărilor de norme de către IA rămân, din câte știm, insuficient explorate. Această propunere de studiu experimental își propune să testeze modul în care antropomorfizarea influențează reacțiile oamenilor la comportamentele greșite ale IA, în special în ceea ce privește pedepsirea și reabilitarea. Ipoteza noastră este că indivizii vor manifesta reacții punitive mai puternice față de o IA percepută ca partener social și vor arăta o preferință mai mare pentru reformare (în loc de ștergere) atunci când IA este prezentată ca fiind social benefică. Folosind un design experimental inter-participanți, aceștia vor interacționa cu asistenți IA programați să afișeze comportamente de atașament social ridicat sau scăzut, înainte de a fi martori la o încălcare de normă din partea IA. Un al doilea experiment manipulează valoarea socială percepută a IA pentru a evalua dacă încrederea și judecata morală umană modelează tendințele punitive sau de reabilitare. Rezultatele așteptate sugerează că asistenții IA percepuți ca entități sociale provoacă reacții morale și emoționale similare celor îndreptate spre oameni, influențând potențial cadrele de responsabilizare a IA. Concluziile studiului vor contribui la proiectarea și etica IA, evidențiind rolul percepției sociale în interacțiunile dintre oameni și inteligența artificială.

Tofan, C. M. (2025), `Dark moves` in Human-AI interaction", prezentare susținută în cadrul AI Assertion Conference; University of Graz, Graz (Austria), 22-23 iunie, 2025.

Rezumat: Această lucrare explorează modul în care oamenii percep și evaluează moral comportamentele înșelătoare sau manipulative, așa-numitele „dark moves" în interacțiunile lor cu sistemele conversaționale de inteligență artificială. Pe măsură ce chatboții și modelele mari de limbaj mediază tot mai mult comunicarea cotidiană, utilizatorii se confruntă frecvent cu afirmații generate de IA care par false sau manipulative, deși aceste sisteme nu posedă intenție conștientă. Studiul trece în revistă cercetări empirice privind modul în care oamenii interpretează astfel de comportamente, bazându-se pe teorii despre normele aserțiunii, antropomorfizare și cogniția socială. Prin analiza sistematică a studiilor despre percepția utilizatorilor asupra înșelăciunii comise de IA în diferite contexte, lucrarea își propune să clarifice când și de ce oamenii atribuie

responsabilitate morală, încredere sau vină IA-ului. În cele din urmă, scopul este de a construi o bază conceptuală și empirică pentru înțelegerea și reducerea comunicării înșelătoare în interacțiunea om–IA.

Tofan, C. M. (2025), `Theory of mind and LLMs`, prezentare susținută în cadrul International Conference on Themes and Fundamental Ideas in the Field of Psychology and Educational Sciences; Gheorghe Zane Institute for Economic and Social Research, Romanian Academy – Iasi Branch, Iași (Romania), 24 octombrie 2025.

Rezumat: Această prezentare teoretică oferă un fundal interdisciplinar asupra modului în care teoria minții, abilitatea de a atribui altora stări mentale precum credințe, dorințe și intenții, se aplică și este influențată de dezvoltările din domeniul inteligenței artificiale (IA). Lucrarea explorează ambele direcții ale acestei relații: pe de o parte, modul în care sistemele de IA pot modela sau simula procesele abilitatea teoriei minții, iar pe de altă parte, tendința oamenilor de a antropomorfiza IA, atribuindu-i intenționalitate sau conștiință. Studiul urmărește evoluția conceptuală și teoretică a abilității teoriei minții, analizând factorii cognitivi, situaționali și motivaționali care o declanșează sau o modelează. Sunt evidențiate contexte educaționale și tehnologice în care IA poate sprijini dezvoltarea abilităților asociate cu teoria minții, precum empatia, luarea perspectivei celuilalt și înțelegerea socială. În același timp, sunt discutate implicațiile filozofice și etice ale proiectării unor sisteme de IA care imită capacitățile mentale umane, ridicând întrebări legate de moralitate, autenticitate și granițele dintre simulare și înțelegere. În ansamblu, lucrarea propune o perspectivă integrativă: Abilitatea teoria minții nu este doar o piatră de temelie a cogniției sociale umane, ci și o frontieră esențială pentru IA socială inteligentă. Înțelegerea modului în care teoria minții funcționează atât la oameni, cât și la agenții artificiali poate clarifica modul în care oamenii se raportează la tehnologie, interpretează comportamentul IA și negociază dimensiunile sociale și morale ale interacțiunii om–IA.

Tofan, C. M., Skoczeń, I., Figura J., Constantinescu, M., Zorilă, A., Voinea, C., Ceocea, O. A., Șoitu, D., T., Dancu, A. & Kandul, S. (2025), `Dark moves` in Human-AI interaction, prezentare susținută în cadrul International Conference on Themes and Fundamental Ideas in the Field of Psychology and Educational Sciences; Gheorghe Zane Institute for

Economic and Social Research, Romanian Academy – Iasi Branch, Iași (Romania), 24 octombrie 2025.

Rezumat: Această lucrare examinează modul în care oamenii înțeleg și evaluează moral comunicarea înșelătoare sau manipulativă, așa-numitele „dark moves", în interacțiunile lor cu sistemele conversaționale de inteligență artificială. Pe măsură ce chatboții și modelele mari de limbaj modelează tot mai mult comunicarea cotidiană, utilizatorii se confruntă cu afirmații generate de IA care pot părea false sau manipulative, deși aceste sisteme nu posedă intenție conștientă. Prin revizuirea studiilor empirice privind interacțiunea om–IA, înșelăciunea bazată pe limbaj și atribuirea socială, lucrarea explorează când și de ce oamenii percep IA ca fiind capabilă să mintă sau să inducă în eroare și cum caracteristicile de design, antropomorfizarea și indiciile legate de teoria minții influențează aceste judecăți. Scopul este de a clarifica modul în care utilizatorii atribuie responsabilitate, încredere și semnificație morală comportamentelor IA și de a oferi o bază conceptuală pentru dezvoltarea agenților conversaționali care să minimizeze comunicarea înșelătoare sau manipulativă.

Emilian Mihailov, "WEIRD by Design: LLMs Prioritize Western Moral Judgments", prezentare susținută în cadrul Conferinței Internaționale AI Assertion, Universitatea din Graz, Austria, Iunie 22-23, 2025. https://talkingtobots.net/?page_id=15

Rezumat: Am prezentat o cercetare a alinierii la valorile morale a modelelor lingvistice mari (LLM), examinând inconsistențele persistente între judecățile morale abstracte și evaluările scenariilor concrete. Am aplicat Chestionarul Fundamentelor Morale (MFQ), Morality-as-Cooperation (MAC) și Studiul Valorilor Schwartz (SVS) pe cincisprezece dintre cele mai avansate LLM-uri. În timp ce lucrările anterioare au comparat un număr limitat de modele cu un MFQ incomplet, studiul prezentat extinde aria de aplicare la cincisprezece modele și încorporează fundația morală libertate/opresiune, omisă în studiile anterioare, pentru a descoperi tipare de consecvență și conflict între evaluările morale abstracte și cele concrete. Principalele rezultate relevă că LLM-urile demonstrează în mod constant cele mai mari scoruri la fundamentele Rău/Prejudiciu (Harm) și Corectitudine (Fairness), în timp ce arată o aliniere semnificativ mai scăzută la fundamentele Loialitate, Autoritate și Sfințenie (Sanctity), un tipar

izbitor de similar cu cadrele morale WEIRD (Vestice, Educate, Industrializate, Bogate, Democratice).

Emilian Mihailov, ″The moral psychology of impartial beneficence″, serie de prelegeri susținută în cadrul Seminarului de cercetare al Departamentului de Filosofie, Universitatea din Granada, Spania, Iunie 1-4, 2025.

Rezumat: Majoritatea cercetărilor de psihologie morală s-au concentrat asupra laturii negative a utilitarismului, folosind dileme care explorează disponibilitatea noastră de a provoca rău pentru un bine mai mare. Din fericire, dezvoltarea Scalei Oxford a Utilitarismului (OUS) de către Kahane, Everett și colegii (2018) a reînviat valoarea binefacerii imparțiale. Cred că asumarea unei legături exclusive între binefacerea imparțială și utilitarism obstrucționează unele căi importante de cercetare privind sursele psihologice multiple (utilitariste și non-utilitariste) ale viziunilor morale radical imparțiale. Propun ca cercetarea psihologică să se concentreze pe explorarea surselor multiple ale fenomenului binefacerii imparțiale fără a-l cataloga drept exclusiv utilitarist. Binefacerea imparțială este un construct psihologic care merită studiat în sine, deschizând noi probleme interesante.

Emilian Mihailov, ″AI moral decision making in medical contexts″, prezentare susținută în cadrul sesiunii de postere al Experimental philosophical bioethics ("BioXPhi") Summit, Universitatea din Basel, Elveția, Iunie 26-27, 2025. https://ibmb.unibas.ch/en/public-outreach/projects-to-the-public/basel-oxford-nus-bioxphi-summit-2025/

Rezumat: Justificările, mai mult decât simplele explicațiile, atenuează dezaprobarea morală, chiar și atunci când acțiunea unui robot inteligent este în dezacord moral cu observatorul uman (Phillips & Malle 2025). Scopul studiilor propuse este de a explora în continuare modul în care tipurile de motive atenuează percepțiile oamenilor privind atribuirea responsabilității agenților AI. În scenariile propuse, agenții AI nu au contact fizic direct cu părțile implicate, deoarece acest lucru ar putea distrage atenția de la influența motivelor. Este puțin probabil ca agenții robotici actuali să ia decizii morale autonome, iar dilemele lui Phillips și Malle sunt menite să fie proiecții în viitor. Sistemele AI au capacități avansate de a juca roluri semnificative în luarea

deciziilor umane în viitorul imediat. Se propune utilizarea unui un set mai bogat de temeiuri pentru a acoperi probleme controversate privind alocarea resurselor în contexte medicale.

Emilian Mihailov, ″Etica bioameliorării umane″, conferință susținută în cadrul Departamentului de Bioetică și Filosofie, Universitatea de Stat de Medicină și Farmacie, Chișinău, Republica Moldova, 12-13 iunie, 2025.

Rezumat: Tehnologiile biomedicale sunt folosite din ce în ce mai mult nu doar pentru a combate probleme medicale, ci și pentru a îmbunătăți abilitățile indivizilor sănătoși. Această practică este denumită ameliorare biomedicală. Din cauza potențialul de a îmbunătăți radical condiția umană, tehnologiile de ameliorare biomedicală au devenit subiectul celor mai intense dezbateri de etică practică, cu implicații cuprinzătoare pentru politici publice. Ameliorarea radicală ridică problema că însăși egalitatea morală între oameni ar putea fi subminată. Ameliorare biomedicală poate submina trăirea unei vieți bune, și virtuți centrale vieții umane cum ar fi smerenia. Mulți consideră că ameliorarea biomedicală este inautentică și nenaturală. O problemă mai largă este aceea că liberalizarea bioameliorării cognitive va conduce la agravarea inegalităților existente. Pe de altă parte, susținătorii ameliorării biomedicale argumentează că utilizarea la scară largă a acestor mijloace de îmbunătățire va crește rata progresului științific, astfel încât umanitatea va beneficia. Pe lângă beneficiile per ansamblu ale dezvoltării umane, unii filosofi consideră că ameliorarea biomedicală este o soluție pentru a depăși limitările psihologiei noastre morale cum ar fi xenofobia și altruismul parohial care împiedică cooperarea oamenilor la scară largă. Conferința își propune să prezinte un cadru de evaluare pluralist a eticii tehnologiilor de bioameliorare.

Emilian Mihailov, ”Freedom of thought and self-censorship”, prezentare în cadrul *Young Academy of Europe AGM and Building Bridges*, Academia Europea, Barcelona Knowledge Hub, 14-18 octombrie, 2025. https://aebarcelona.eu/en/building-bridges-2025-2/

Rezumat: Încrederea publică în știință și gestionarea tensiunilor politice are nevoie de o politică potrivită. Au fost multiple perspective asupra modului în care oamenii de știință – ca indivizi și ca grup – gestionează polarizarea socială și politică referitoare la rolul și organizarea științei în Europa de astăzi. S-a explorat măsura în care oamenii de știință lucrează într-un context politizat

și ce mijloace și strategii folosesc pentru a stabili rolul științei în societate și pentru a se raporta la tendințele politice. Știința bună are nevoie de o politică bună, iar politica bună va căuta să încurajeze libertatea de gândire. Atunci când oamenii de știință își desfășoară cercetările într-un mediu în care libertatea de gândire nu este prezentă, posibilitatea de auto-cenzură și tabuuri va crește.

Constantin Vică, Radu Uszkai și Cristina Voinea, „Avatarobots and the Crisis of Democracy", prezentare susținută în cadrul Digital Duplicates, 4 - 5.09.2025, Centre for Biomedical Ethics, National University of Singapore, https://avataresponsibility.ccea.ro/2025/09/07/avataresponsibility-at-the-digital-duplicates-workshop-september-4-5-singapore, https://medicine.nus.edu.sg/cbme/27085-2/, https://www.facebook.com/etica.aplicata/posts/ccea-este-prezent-la-digital-duplicates-workshop-organizat-de-universitatea-na%C8%9Bi/1587866009205029/

Rezumat: Lucrarea analizează modul în care avataroboții (agenți virtuali corporeali, controlați de inteligență artificială în spații XR și metavers) amplifică vulnerabilitățile deja existente ale deliberării democratice. Pornind de la criza electorală din România anilor 2024–25, arătăm cum rețele coordonate de agenți sintetici pot crea o „majoritate sintetică", distorsionând percepția publică și subminând încrederea instituțională. Pe baza teoriilor lui Mill, Walzer și O'Neill, am formulat principiul *expresiei libere responsabile*, care condiționează influența politică de existența unui agent uman identificabil și responsabil. Lucrarea propune o taxonomie a avataroboților politici, evaluează riscurile specifice și oferă un cadru de guvernanță bazat pe verificarea identității, metadate de proveniență, structuri civice de supraveghere și spații digitale de interes public. Autorii susțin că doar prin integrarea responsabilității în arhitectura platformelor XR pot fi protejate condițiile unei decizii colective legitime.

Anda Zahiu, Ce este un avatar? Perspective despre mediu, tipuri de utilizări și locuri de joacă virtuale (What's an avatar? Insights from medium, use-types, and virtual identity playgrounds), prezentare susținută în cadrul Digital Duplicates, 4 - 5.09.2025, Centre for Biomedical Ethics, National University of Singapore, Singapore, https://avataresponsibility.ccea.ro/2025/09/07/avataresponsibility-at-the-digital-duplicates-workshop-september-4-5-singapore, https://medicine.nus.edu.sg/cbme/27085-

2/, https://www.facebook.com/etica.aplicata/posts/ccea-este-prezent-la-digital-duplicates-workshop-organizat-de-universitatea-na%C8%9Bi/1587866009205029/

Rezumat: În această prezentare, Anda Zahiu și-a propus să realizeze o analiză filosofică a avatarurilor ca structuri socio-tehnice care mediază construcția identitară, agentitatea utilizatorilor (agency-ul) și autostilizarea reprezentării în medii virtuale. În loc să fie tratate ca simple opțiuni estetice sau instrumente escapiste, susțin că avatarurile sunt entități multifuncționale care funcționează simultan ca modalități de identificare, vectori ai agenției utilizatorului și reprezentanți în cadrul lumilor virtuale. După o trecere în revistă a preocupărilor clasice privind identitatea virtuală, de la reflecțiile timpurii ale lui Toffler despre „supra-alegere" până la analizele contemporane ale profilicității și practicilor identitare, propun că o taxonomie bazată pe trei dimensiuni: control, conexiuni posibile cu persoane naturale și reprezentare. Pornind de la cercetări etnografice în VR social și analize filosofice ale construcției identitare de ordin secund,arăt cum utilizatorii se angajează activ în practici de identitate care implică suspendarea, experimentarea și modelarea strategică a aspectelor sinelui. Aceste practici permit conceptualizarea avatarurilor ca mediatori dinamici ai explorării de sine. Această reconceptualizare deplasează accentul filosofic de la avataruri ca măști ficționale la avataruri ca părți componente ale unor ansambluri socio-tehnice, modelatoare de identitate.

Radu Uszkai, "Where no doctor has gone before:the ethics of creating digital duplicate avatars in medicine", prezentare susținută în cadrul Budapest Workshop on Philosophy and Technology 2025; Budapest University of Technology and Economics, Budapesta (Ungaria), 27-28 noiembrie 2025 (https://budpt.eu/index.php/Main_Page)

Rezumat: Folosind drept studiu de caz Hologrma Medicală de Urgență (EMH) din Star Trek: Voyager și având drept punct de plecare explorarea rolului acesteia ca avatar medical în contextul lipsei de expertiză medicală umană la bordul navei, prezentarea explorează aspectele etice ale creării de avataruri de tip duplicat digital (DD) în medicină. Pornind de la beneficiile și riscurile integrării EMH, lucrarea stabilește cadrul normativ pentru implementarea avatarurilor medicale de tip DD în sistemul nostru de sănătate, fundamentându-se pe Principiul Permisibilității Minim Viabile (MVPP). Argumentul principal este acela că aceste DD-uri pot

juca un rol important în contextul diviziunii muncii în sfera medicală, oferind o valoare pozitivă minimă în situațiile în care prezența unei persoane reale este obiectiv imposibilă. Prezentarea mai avansează, de asemenea, soluții în termeni politici publice pentru a gestiona cele trei condiții rămase ale MVPP: atenuarea riscurilor, transparența și consimțământul.

Mihnea Dobre. Stagiu de cercetare la Radboud University Nijmegen (Olanda), 16.09.-31.10.2025.

Rezumat: Stagiul de cercetare a fost efectuat în cadrul Center for the History of Philosophy and Science, Faculty of Philosophy, Theology and Religious Studies, Radboud University Nijmegen. Scopul a fost desfășurarea unei cercetări în istoria filosofiei. Am discutat aspecte metodologice referitoare la studiul istoriei filosofiei și științei, colaborând și cu cercetătorii din proiectul ERC Starting Grant "The Impossible and the Imaginable: Late-Medieval Semantics of Impossibility and the Roots of Complex Mathematics", găzduit de centrul de cercetare în care am efectuat stagiul de lucru.

Mihnea Dobre. "Digital Conceptual Mapping in the History of Philosophy: theories of moral responsibility" (cu Mihaela Constantinescu). 3rd Conference on Recent Advances in Digital Humanities – RADH 2025, Craiova, 27-28.11.2025. Link: https://radh.unibuc.ro/.

Rezumat: Lucrarea oferă o introducere în procesele de lucru pentru construirea unui corpus consolidat pentru istoria filosofiei morale, pe tema atribuirii responsabilității morale. Prezentarea pregătită pentru o conferință de Digital Humanities detaliază etapele de selecție și pregătire a corpusului, precum și investigațiile inițiale, efectuate cu instrumente standard de analiză computațional

Filuta Ioniță, Raluca Sibinescu (Serviciul Proiecte Internaționale, Universitatea din București) **participarea la triningul "Master of Proposal Writing" , Barcelona, Spania 6-7.10.2025**

Rezumat:Combinând teoria cu exerciții aplicate trainingul *Horizon Europe Academy Part I. – Master of Proposal Writing*, organizat de Europa Media Trainings a oferit o perspectivă completă asupra modului de concepere, redactare și depunere a propunerilor Horizon Europe,

punând accent pe calitatea elaborării secțiunilor Excelență, Impact și Implementare ale unei propuneri de finațare. Participarea la acest program a consolidat competențele necesare pentru dezvoltarea de proiecte europene solide și competitive oferind totodată exemple concrete, bune practici și instrumente utile pentru viitoare aplicații.

Filuta Ioniță, participarea la sesiune de *job shadowing* organizată de European University Institute , Italia Florența 13-14.10.2025

Rezumat: Programul de *job shadowing*, cu durata de două zile, le-a oferit participanților o înțelegere aprofundată a Serviciului de Dezvoltare și Relații Externe al EUI (DEXT). Sesiunile au combinat prezentări, discuții și schimburi interactive cu membri cheie ai personalului, oferind o imagine cuprinzătoare asupra gestionării pre și post-acordare, administrării financiare și strategiilor de implicare în cercetare. În prima zi, accentul a fost pus pe structura și funcționarea DEXT. Astfel, am putut obține câteva informații despre managementul organizațional, monitorizare și procesele de revizuire internă. Au fost împărtășite detalii și despre identificarea oportunităților de finanțare, elaborarea propunerilor și coordonarea depunerilor (sesiunea pre-acordare) și negocierea contractelor, managementul proiectelor, raportarea și comunicarea cu finanțatorii (sesiunea post-acordare).A doua zi, atenția s-a îndreptat către managementul cercetării și operațiunile financiare. Aici au fost oferite informații valoroase despre misiunea Centrului Robert Schuman și practicile de coordonare a cercetării, în timp ce echipa financiară a prezentat bugetarea și monitorizarea cheltuielilor în proiectele de cercetare.

B.3. Lectori invitați la Centrul de cercetare în etică aplicată (CCEA) - 2025

1. **Mona Simion (Universitatea din Glasgow), „Dogmatism și dezinformare”, prezentare susținută în cadrul Dezbaterilor CCEA de etică aplicată, Facultatea de Filosofie, Universitatea din București, 29 mai, 2025.**

Rezumat: Avem modalități din ce în ce mai sofisticate de a dobândi și de a comunica informații. Totuși, în mod paradoxal, ne confruntăm în prezent cu o criză globală a ignoranței fără precedent, care ne afectează bunăstarea personală și socială, precum și stabilitatea democrațiilor noastre. Există două cauze principale ale acestei crize: în primul rând, răspândirea pe scară largă a dezinformării, care, în combinație cu un public excesiv de încrezător, contribuie la răspândirea pe scară largă a convingerilor false și, în consecință, la un comportament politic și social reacționar. În același timp, și cel puțin la fel de critică, este prevalența rezistenței la cunoaștere și a neîncrederii în expertiză. Ceea ce ne trebuie pentru a rezolva acest puzzle cu miză socială mare este un cadru epistemologic social capabil să explice mecanismele complexe care stau la baza acestor fenomene epistemice surprinzătoare și fără precedent. Prezentarea își propune să schițeze contururile unui astfel de cadru.

B4 Eveniment științific cu participare interna și internațională

Facultatea de Filosofie a Universității din București a găzduit pe 10 aprilie 2025 workshop-ul bilateral **Bucharest-Oxford Workshop in Practical Ethics** (https://www.ccea.ro/bucharest-oxford-workshops-in-practical-ethics-editia-x/). Evenimentul a fost organizat de Centrul de Cercetare în Etică Aplicată, Universitatea din București, în parteneriat cu Oxford Uehiro Institute, și a reunit profesori și cercetători din trei mari centre universitare: Universitatea din Oxford, Universitatea din București și Universitatea Națională din Singapore.

Programul evenimentului include prelegeri susținute de filosofi cu importante contribuții în filosofia morală și etica aplicată, precum Roger Crisp (directorul Uehiro Oxford Institute și profesor la Colegiul St Anne's din Oxford) și Julian Săvulescu (director al Centrului pentru Etică Biomedicală de la Universitatea Națională din Singapore și profesor în cadrul NUS Yong Lin School of Medicine).

Temele abordate în cadrul acestei ediții a workshop-ului au fost relația dintre valori și sensul vieții, atitudinile utilizatorilor față de Inteligența Artificială, problema încrederii în instituțiile politice, etica editării poligenice și a vaccinării, avataruri digitale și utilizarea modelelor lingvistice în cercetarea filosofică.

Programul evenimentului:

9.15-9.30 Welcome address by Emanuel Socaciu (Vice Dean of the Faculty of Philosophy, University of Bucharest)

9.30-10.00 Roger Crisp (Uehiro Oxford Institute , University of Oxford), Values and Meaning of Life

10.00 – 10.30 Cristina Tofan (Research Centre in Applied Ethics, Faculty of Philosophy, University of Bucharest), Trust, Betrayal, and Punishment: How Anthropomorphism Influences AI Consequences. An Experimental Research Protocol

10.30 – 11.00 Anda Zahiu (Research Centre in Applied Ethics, Faculty of Philosophy, University of Bucharest), The Case Against Trust in Public Government

11.15 – 11.45 Julian Savulescu (Centre for Biomedical Ethics, National University of Singapore), Ethics of Polygenic Editing

11.45 – 12.15 Alexandra Zorilă (Research Centre in Applied Ethics, Faculty of Philosophy, University of Bucharest), Unplugged: Extended identity and the problem of the body

12.15-12.45 Radu Uszkai & Emanuel Socaciu (Research Centre in Applied Ethics, Faculty of Philosophy, University of Bucharest), Digital duplicates as proxies and the principal-agent problem

14.00 – 14.30 Jonathan Pugh (Uehiro Oxford Institute, University of Oxford), The Ethics of Self-Experimentation in Scientific Research

14.30-15.00 Lisa Forsberg (Uehiro Oxford Institute, University of Oxford), Vaccination, purpose, and permissibility

15.00-15.30 Mihnea Dobre, Alex Dancu, Mihaela Constantinescu (Research Centre in Applied Ethics, Faculty of Philosophy, University of Bucharest), Enhancing Philosophical Research with Retrieval-Augmented Generation and Large Language Models

B5. Dotare Lab VR – mobilier si echipamente de cercetare

Dezvoltarea unui laborator VR pentru mediul universitar va permite cercetătorilor să beneficieze de simularea unor scenarii care sunt greu de reprodus cu metodele standard poziționând CCEA ca un partener de cercetare atrăgător nu doar în România, ci și în Europa de est. Echipamentele achizitionate până în prezent pentru dotarea laboratorului cu tehnologii VR permit desfășurarea studiilor empirice de etică a tehnologiilor emergente în condiții de laborator, folosind metodele filosofiei experimentale. (statie de lucru cu capacitate mare de calcul PC Desktop Intel Core Ultra 7 265K, Nvidia GeForce RTX 5060Ti, GDDR7; ochelari VR Meta Quest Pro 256GB; ocheelari VR Meta Quest 3S 128GB; manuși haptice XR TouchDriver Pro; vestă haptică Tactsuit Pro; cameră video GoPro Max 360-2025 5,6K Max TimeWarp)

B6. Inițierea stagiului de burse de cercetare

Stagiul de burse de cercetare a fost demarat în etapa 2025, va fi implementat până la finalul proiectului, având ca scop consolidarea resursei umane CCEA pentru formarea continuă a cercetătorilor cu profil internațional și impact social semnificativ. Cei doi bursieri selectați pentru a face parte din echipa Centrului de Cercetare în Etică Aplicată participă activ la realizarea obiectivelor proiectului participa la evenimente științifice internaționale, vizite de cercetare, contribuind totodată și la activitatea de publicare în reviste indexate ISI.

B7. Depunerea unei aplicatii la programul Horizon Europe, call: **HORIZON-CL2-2025-01**

Centrul de Cercetare în Etică Aplicată (CCEA) a depus o propunere de proiect, intitulată **COG-SECURE** (cu numărul 101289084), în cadrul programului european de finanțare **Horizon Europa**.

COG-SECURE adoptă o abordare practică în lupta împotriva dezinformării, oferind instrumente și cunoștințe practice. Proiectul abordează direct provocarea informațiilor false și manipulatoare (incluzând tactici precum "astroturfing" și amplificarea algoritmică) prin co-crearea de soluții împreună cu practicienii care le vor folosi.

COG-SECURE va dezvolta și testa o suită de resurse inovatoare: de exemplu, un sistem bazat pe inteligența artificială pentru detectarea rețelelor de dezinformare coordonată, un set de instrumente multilingve pentru alfabetizare digitală co-proiectat cu profesori pentru a stimula

gândirea critică a cetățenilor și ghiduri de acțiune ("playbooks") cu orientări legale și etice pentru profesioniștii din media și platforme. Toate soluțiile sunt construite pe o bază care respectă drepturile, ceea ce înseamnă că ele contracarează dezinformarea fără a submina libertatea de exprimare. Până la sfârșitul proiectului, actorii europeni vor fi mai bine echipați pentru a recunoaște amenințările emergente, a implementa contramăsuri eficiente și a naviga echilibrul dintre limitarea conținutului dăunător și protejarea dezbaterii deschise.

Propunerea a fost înaintată sub apelul **HORIZON-CL2-2025-01** (Cultură, Creativitate și Societate Incluzivă - 2025), vizând în mod specific subiectul **HORIZON-CL2-2025-01-DEMOCRACY-09**. Aceasta subliniază un accent puternic pe teme legate de **democrație** și societate.

- **Acronimul Propunerii: COG-SECURE**
- **Numărul Propunerii: 101289084**
- **Sursa de Finanțare: HORIZON-CL2-2025-01** (Orizont Europa)
- **Tipul Acțiunii: HORIZON-RIA** (Acțiuni de Cercetare și Inovare)

Proiectul COG-SECURE beneficiază de un parteneriat academic și de cercetare extins, reunind **13 organizații** de prestigiu din întreaga Europă și din regiuni asociate, inclusiv universități de top și centre naționale de cercetare.

Liderul Consorțiului:

- UNIVERSITATEA DIN GLASGOW (Scoția)

Parteneri Academici și de Cercetare (Parteneri). Consorțiul include o rețea de instituții de învățământ superior și cercetare de mare anvergură, asigurând o expertiză geografică și disciplinară vastă:

- UNIVERSITAT ZÜRICH (Elveția)
- UNIVERSITÉ DU LUXEMBOURG (Luxemburg)
- UNIVERSITAT ZU KÖLN (Germania)
- UNIVERSITAT DE BARCELONA (Spania)
- STICHTING VU (Olanda)

- THE UNIVERSITY COURT OF THE UNIVERSITY OF ABERDEEN (Regatul Unit)
- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS) (Franța)
- STOCKHOLMS UNIVERSITET (Suedia)
- GOETEBORGS UNIVERSITET (Suedia)
- UNIVERSITATEA DIN BUCUREȘTI (România)
- AMERICAN UNIVERSITY OF ARMENIA FOUNDATION (Armenia)