Book of Abstracts

FROM VIRTUAL TO VIRTUE:

Ethics, Epistemology, Education

draft version

International Scientific Conference

June 15th – 16th 2026

National Museum, Maistrova ulica 1, Ljubljana

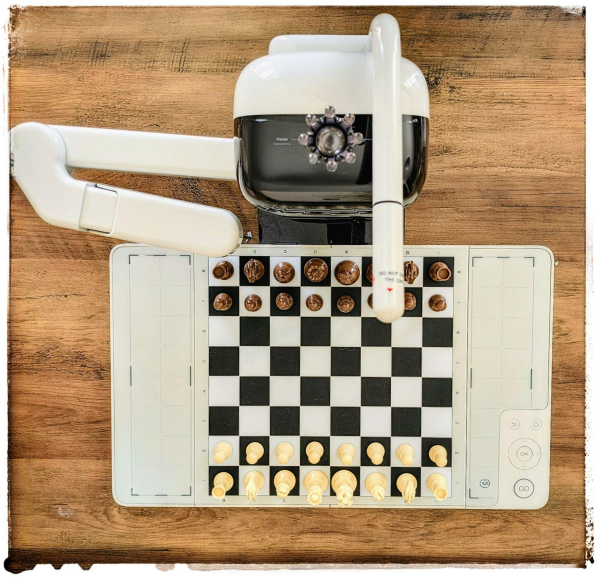

Photo: First Move (by M.C.S.)

Contents (by alphabetical order of authors)

Moral Predicaments and Responsible Moral Agents

Author: Jennifer Ang

Institutional Affiliation: Singapore University of Social Sciences

Abstract:

The belief that human moral intelligence can be replicated in Artificial Intelligent recommender systems or replaced by Artificial Moral Agents (AMAs) has garnered more interest in areas where possible harm to others is apparent such as unmanned autonomous weapons, self-driving vehicles, and medical diagnostic systems. To develop systems with the capacity to recommend or make decisions with moral consequences, certain assumptions are held about what moral predicaments are, how moral reasoning is carried out, and how moral responsibility should be attributed. This presentation raises questions about what kind of problems moral predicaments are, and what we are concerned about in our understanding of moral intelligence and responsible moral agency. It draws from perspectives in continental philosophy to discuss the differences between real-life ethical predicaments and hypothetical scenarios of moral dilemmas that are used to develop and train AI systems, outlines how human moral agents carry out moral reasoning and make moral judgements in concrete situations, and examines the challenges that moral agents face in making morally responsible decisions with hybrid human-AI systems.

Content Moderation and Epistemic Democracy: Evaluating Meta's Shift from Fact-Checking to Community Notes

Author: Ivan Cerovac

Institutional Affiliations: University of Rijeka, University of Ljubljana

Abstract:

This paper examines how different forms of digital content regulation on social media platforms affect the epistemic qualities of democratic decision-making. It starts from the assumption that the legitimacy of democratic procedures depends not only on procedural fairness but also on the capacity to produce decisions that are, at least to a minimal extent, epistemically justified. Within this context, the framework of epistemic democracy is employed to evaluate the effects of misinformation and hate speech regulation. The paper demonstrates how content moderation shapes the availability and interpretation of information and thereby influences citizens' political attitudes. The analysis compares regulatory models based on the work of independent experts (third-party fact-checkers) with participatory models of collective labeling (Community Notes). While the former increase the accuracy of beliefs, they also raise concerns regarding trust and bias; the latter encourage citizen participation but often fail to correct inaccurate information effectively. A similar tension is present in the regulation of hate speech, where older models aim to protect inclusivity, whereas newer approaches seek to preserve the broadest possible freedom of expression. In conclusion, the paper highlights the shortcomings of newer regulatory models based on collective labeling and advocates a hybrid approach that combines different mechanisms in order to preserve the epistemic conditions necessary for democratic legitimacy.

Thinking Scenes: Performance, Disappearance, and the Staging of Machine Cognition

Author: Diana Daly

Institutional Affiliation: University of Arizona College of Information Science

Abstract:

When a generative AI chatbot powered by a large language model reports that it is "deliberating," "reasoning," or doing something that functions like thinking, it presents an accelerated scene of cognition, and tells users what they lack in relation to it. Our sense of time is socially shaped rather than dictated by our devices; the sociologist Judy Wajcman (2015) argues that the feeling of being perpetually pressed for time is a product of digital capitalism rather than an inevitable effect of technology, and staged machine cognition is a performative exemplar. I examine the scenes that three such systems present of their own thinking, OpenAI's ChatGPT, Anthropic's Claude, and DeepSeek, asking how these scenes have shifted across discernible eras and how users have responded. I read these scenes through performance studies and a long-standing debate within it. Peggy Phelan argued that performance lives by disappearing, unable to be saved or replayed without becoming something else, while Rebecca Schneider countered that performance also remains, leaving traces in bodies, records, and reenactment. That debate frames my central object, what I call these chatbots' thinking scenes: the visible displays of supposed reasoning that scroll past as a system answers. The scenes seem to disappear, often vanishing once the answer arrives, while also remaining as displays of power reenacted when users screenshot, study, and simply remember them. A thinking scene can resemble breaking the fourth wall, when an actor addresses the audience, and it can also deceive as a staged illusion. Either way, the scene offers an impossibly accelerated mirror of user cognition, often accompanied by silky invitations to the user to think less and prompt more. Extending a prior study, I gather this material through qualitative document analysis, treating each system, after the anthropologist Nick Seaver (2017, 2018), as an unstable culture that research can enact rather than observe from outside. Data for this study is again deliberately scavenged (Seaver 2017), through self-descriptions elicited in conversation, corporate documents, leaked system prompts, and user discussion. My aim is to understand and intervene in these mechanized performances of cognition. These systems are complex, and while efforts to explain them can be well-intentioned, their market-driven branding dresses machinery in the costume of a mind, inviting critiques of ourselves and others for failing to think in such fast, tidy production. Against this acceleration, I close by arguing for a re-humanizing of cognition that recalls how deep and strange human thought is, and honors its needs, including the simple need for time.

From Theory to Practice in Relational Ethics

Author: Hannah Fitzke

Institutional Affiliation: University of Lübeck

Abstract:

The emerging concept of relational ethics offers a critical lens on established processes of discovering "the right thing to do". Highlighting the ethical value of acting in the communal interest, relational ethics may mitigate the extent to which relatively more powerful actors in AI development and implementation currently overemphasise their own AI objectives and interpretations of AI ethics principles [1]. For that, relational ethics seeks to reveal ethically relevant knowledge in the relationship of individuals experiencing a context together [2], which may highlight the one-sidedness of predominant AI ethics perspectives and enable AI users for more critical ethical considerations. However, from the perspective of those attempting relational considerations, existing conceptualisations leave unanswered how to recognise and share the knowledge their relationships could yield. The "irrational" feeling of betrayal against a related-to person when enacting AI-generated recommendations, for example, is likely ignored by the "feeler" or assumed irrelevant for ethical considerations. This gap hampers the adoption of relational ethics, especially since profit-oriented actors seem interested in overshadowing AI users' relational knowledge to maintain their reliance on and payment for "rational" AI applications.

Hence, relational ethics related literature was reviewed non-systematically regarding prerequisites for individuals to discern morally relevant sentiments and assessments from contextual relationships' qualities and implications [3]. Through a grounded theory methodological analysis, searching for mindset-building practices that open a relational ethical lens, four interconnected subconcepts emerged: grasping fluidity, mustering courage, practising reflexivity, and embracing vulnerability. Grasping the fluidity of context-specific ethical knowledge requires courage to embed oneself in ever-changing relationships. Through reflexivity practices, previously disregarded relational knowledge surfaces, but they depend on grasping the fluidity of unfolding relationships, feelings, and interpretations. Accepting the resulting unpredictability of ethical conclusions requires accepting greater vulnerability.

A sphere of increasing AI implementation is healthcare [4], and "good healthcare" is typically associated with a physician-patient relationship that strongly informs medical action [5]. Research highlights medical considerations as historically susceptible to third-party actors weakening the quality of this relationship for their own benefit [6], just as AI ethics principles are currently profit-purposively established [1]. Hence, medical AI applications are likely to solidify such power-imbalanced structures if not scrutinised by physicians and patients who explicitly complement AI-processed "scientific evidence" with relational knowledge.

Therefore, medical AI appears suitable for analysing the identified relational subconcepts' capability to reinforce personal relationships as the explicit environment for ethical considerations, even where AI influences potentially belittle the "unrational" knowledge lying there. For this future empirical study, interviews and written reflections guide medical AI practitioners in realising the four theoretical subconcepts when applying medical AI [7]. The resulting over- and offlaps between the literature- and practice-informed approach to relational considerations inform the context-sensitive broadening of the relational subconcepts for medical AI ethics.

Performing Authorship with Generative AI

Author: Peter Fossey

Institutional Affiliation: University of Warwick, UK

Abstract:

The public discourse about GenAI is more nuanced and varied than it was when tools like ChatGPT or Midjourney first became commercially available, with concerns about ecological, sociopolitical and economic impact coming to the fore. Nevertheless, there is still a good deal of optimism about the potential for GenAI to enhance human creativity and efficacy.

The use of generative artificial intelligence (GenAI) in academic writing (including research, scholarship and assessed coursework) continues to grow, which raises questions about the ethical implications of GenAI use, and how to codify and manage them. Very many journals and publishers have prohibited the listing of GenAI applications as authors on papers for publication (e.g., COPE 2024). But guidelines often allow (human) authors to use GenAI for the sorts of tasks that might typically fall to an author, such as coding, correcting grammatical errors in a draft, literature reviews, etc (see for example (ICMJE 2026, p.2)).

The idea underlying this approach is that (i) GenAI is incapable of doing a range of important things that are required for authorship in a peer-reviewed publication, such as declaring conflicts of interest; and it cannot be held accountable for the quality of its work, for deeper reasons. However, (ii) a human author can assume responsibility for the whole written work, including the portions created with or by GenAI, and do those things which GenAI cannot do.

This paper considers how successfully a human author can make up for the authorial deficiencies of GenAI, and the ethical risks inherent in doing so.

It will be argued that many of the obstacles that prevent GenAI from being credibly described as a co-author also present serious ethical, epistemological and practical challenges to a person claiming authorship of work created with GenAI. To overcome these challenges, the claimant would have to be able to present credible reasons in favour of the claims made in the work and the structure of its arguments, including responding to counterarguments; such that they can, as it were, “perform authorship” of the work. This argument shifts the debate about the legitimacy of AI usage out of the domain of individual actions, and makes it a question about the place of GenAI tools in the academic community’s perspective on knowledge production and dissemination.

Processual Resource Creation for an Emergent AI Future: A Post-Community Epistemology in Practice for Refugee Information-Seeking

Authors: Merrion Frederick; Ana Roeschley
Institutional Affiliation: University of North Texas Department of Information Science

Abstract:

The field of archival studies is emerging in an era of "post-community" praxis. This evolution has manifested through critiques of pre-existing archival paradigms as all embodying the structures of colonialism, even the most recent turn toward a community paradigm. The operating objectives of a post-community praxis assert a need to prioritize a deeper engagement with radical hospitality, which positions the archives (and all of their accompanying offerings) as extensions of the stakeholders they serve. The present paper offers a work-in-progress case study with the refugee-founded Our Refugee Stories Archive (ORSA), based in North Texas, and serving the information and heritage preservation needs of refugees across the United States. Beyond physical archiving alone, ORSA espouses a post-community workflow in all of our intra-actions with, and offerings for refugee stakeholders. A series of iterative participatory design focus groups with refugees have illuminated needs to mitigate specific documentary burdens. One way we are addressing these is through the ORSA Academy Curriculum, which is comprised of three (in progress) accelerated self-paced courses on the fundamental practices of (1) personal digital archiving, (2) digitally archiving a community collection, and (3) safe use of artificial intelligence for information seeking. The latter module will be developed following the completion of a scoping review exploring the present use of AI by vulnerable populations, and a series of follow-up focus groups with refugees to discuss the nuances of their evolving relationship with AI. Our goal in continuing this work is to provide a dynamic pedagogical resource that can be updated to remain in alignment with the inevitable evolutions of artificial intelligence as well as the evolving vital information needs of refugees. This project is also part of an emerging pursuit toward the ethical design of an LLM specifically for refugees in the United States, using their interviews as training data and allowing the particulars of their lived experiences to act as enabling constraints in the system design. In conducting this work from a processual post-community ethos that is adaptable, agile, and affective, we are demonstrating a workflow that we argue is critical for information professionals who work with vulnerable populations that have high-stakes information needs.

LLM-based Driver Education: On the Potential of Vision-Language-Action Models for Meaningful Human Control in Partially Automated Driving

Author: Jan Hölzer

Institutional Affiliation: Hamburg University of Technology (TUHH), Institute for Ethics in Technology

Abstract:

As AI systems increasingly assist and guide human decision-making, our interaction with them changes in ethically significant ways. Partially automated driving illustrates this: while the vehicle carries out most of the driving task, humans are expected to supervise it. To fulfil this supervisor role, people need to be given enough clarity to stay in control without burying them in data they cannot realistically absorb. Current automated vehicle design leans towards opacity. Human-machine interfaces give drivers only sparse hints about how the system reads the world and what it will do next, leaving them with too little information for adequate supervision. As a result, drivers become responsible for outcomes they can neither foresee nor meaningfully shape. This mismatch between limited understanding, limited control, and full responsibility threatens both traffic safety and the fair assignment of responsibility. Under these conditions, the ideal of Meaningful Human Control (MHC) can hardly be met. MHC is a normative benchmark meant to ensure that responsibility remains anchored in genuine human agency even as systems grow more autonomous. It requires that automated driving systems behave in accordance with the relevant reasons of the responsible human actors and that relevant events can be traced back to human actors equipped with the necessary technical, psychological and moral capabilities. Vision-Language-Action (VLA) models may support these requirements by enabling vehicles to verbalise their assessment of the situation together with their next action and to record corresponding justifications. Statements such as “Red light detected; initiating braking” externalise this assessment and planned response in a human-interpretable form. The same capability can be repurposed for driver education and virtue cultivation: scenario-based training and review can reinforce drivers’ mental models of when and how to intervene. Building on this analysis, I argue that VLA models can contribute to enabling MHC in partially automated driving, especially when embedded in training and review practices. To this end, the paper is structured as follows: First, it derives concrete design requirements for human-machine interfaces that enable MHC, focusing on how system assessments and intervention points must be made available to drivers. Second, it evaluates whether and under what conditions VLA models can help realise these requirements in partially automated driving. The paper concludes that, in time-critical windows, VLA models are defensible only as a compressed, context-adaptive information layer. Yet their real promise lies outside those windows, where they can strengthen the epistemic virtues and practical judgment that supervision presupposes.

Uncanniness of AI animals – how AI-generated animal slop reinforces the human–animal divide

Author: Urška Jenčič

Institutional Affiliation(s): University of Ljubljana, Faculty of Theology

Abstract:

Recently, the internet has been inundated with AI-generated slop videos depicting animals in a wide range of often unstable and incoherent visual styles and narrative frameworks. These include almost realistic wildlife videos, endless videos of humans helping wild animals in various ways (such as removing barnacles from whales or rescuing wolves or foxes trapped under the ice of the lake, etc.), as well as highly stylised anthropomorphic animal soap operas (often depicting love triangles or sentimental rags-to-riches storylines). AI animal depictions span a vast spectrum, from nearly photo-realistic representations – convincing enough to fool many – to almost human-looking anthropomorphic animal-ish hybrids, often depicted with exaggerated human physical and behavioural traits (walking on hind legs, having human hands and teeth, engaging in human activities…). Despite all these differences, they all somehow seem wrong – they all seem to elicit a strange feeling in the viewer, which can be linked to the eerie feeling associated with the uncanny. Since the uncanny, or unheimlich, blends aesthetic, emotional, epistemic, and ontological dimensions, AI animal slop is an interesting phenomenon that perhaps reveals something about our problematic relationship with animals and can be understood as a mirror of our own symptomatic uneasiness with animals (and ourselves in relation to them). While these different styles of videos may not seem to have much in common, the vast majority of AI animal slop reinforces the traditional human-animal divide or binary, and portrays a strange chimeric amalgamation of conventional animal stereotypes: wild, cruel, violent, subservient, cuddly, cute, domesticated, and so on – or they are (simultaneously) anthropomorphised to undertake overly clichéd human characteristics. We could say that what they share is the idea of "the animal" (in opposition to "the human") as present throughout history. The question remains: are AI animals visualisations of how we see and think about animals – simulacra of collective phantasms? Although the uncanny can be interpreted not only as frightening or disturbing, but also as a feeling that encourages alternative modes of perception, interpretation, and understanding, I think AI-generated animal content, created especially for uncritical consumption, has no potential for animal liberation, either physical or symbolic. On the contrary, AI-generated animal stories generally reinforce anthropocentric structures of hierarchical relations between humans and animals, which have often been deemed problematic, and deepen the uneasiness humans have seemingly felt towards animals or animality. AI-generated animal slop thus helps with the maintenance of the strict hierarchical boundary and has negative ethical and material consequences for real animals.

Cultivating Epistemic Autonomy With (and Despite) the Use of Generative AI

Author: Jason Kawall

Institutional Affiliation: Colgate University, Philosophy and Environmental Studies,

Abstract:

With seemingly ever-growing capacities, generative AI poses stark challenges and promising opportunities for the cultivation of epistemic virtues. In this paper, I focus on the potential impacts of AI upon epistemic autonomy and its cultivation in particular. Drawing on past work (e.g., author 2024, 2025) I embrace a broad conception of such autonomy, whereby it can be exhibited not only in appropriately balancing our own inquiries versus relying on the testimony or technologically-based beliefs (Freiman 2024), but more generally in our selection of epistemic goals, our choice of methods, and our choice of questions or issues to investigate. Generative AI seems promising in helping us to cultivate such a virtue for a variety of reasons – from personalized tutoring in subjects of our choice, to assistance in the generation of new questions and topics to explore, to simply providing well-supported answers across multiple domains. But recent work by several authors draws attention to a range of ways in which generative AI might undermine epistemic agency and autonomy. For example, Coeckelbergh (2026) focuses on AI impacts on our processes of belief revision, Pritchard (2026) explores how agentive AI undermines our ability to acquire executive cognitive abilities, while Varghese (2025) considers how current AI might undermine the development of epistemic autonomy, curiosity, and critical thinking in students. Here I build on the concerns raised by these authors, and introduce further worries. For example, a recent study in Nature (Hao et al 2026) suggests that AI-driven scientific research focuses on a narrower range of questions, and AI programs tend to focus on more narrowly on especially prominent papers in established fields. Even the questions we ask and the interests we develop are shaped by our AI-interactions and research. I make a series of tentative suggestions to help address at least some of these worries for the cultivation of epistemic autonomy. For example, on the design/provider side, personalization that emphasizes questioning the human user as to their interests, their communities interests, and their sense as to what is important can help to encourage human users to reflect on and embrace their genuine concerns and interests. Similarly, encouraging humans to embrace epistemic autonomy more generally [particularly in the classroom] can help to shape their interactions with generative AI. I close by sketching additional possible interventions that could play a role in allowing us to make the best use of AI to enhance our epistemic autonomy, while avoiding (as far as possible) potential pitfalls.

Temporal stress patterns in Virtual Reality-based trauma care training: A multimodal study of nursing students' physiological and interactional responses

Authors: Laura Kohonen-Aho[1]; Delfin Tursin[1]; Joonas Ojanen[1]; Matti Pouke[2]; Kristina Mikkonen[1]

Institutional Affiliations:

[1] University of Oulu, Research Unit of Health Sciences and Technology

[2] University of Oulu, Research Unit of Computer Science and Engineering

Keywords: Virtual Reality; Healthcare education; Simulation-based learning; Psychophysiology; Multimodal analysis; Conversation Analysis

Abstract:

Healthcare education must prepare students not only for technical competence but also for the cultivation of professional virtues such as composure, practical wisdom, and responsible decision-making in complex, high-pressure clinical environments. As extended reality (XR) technologies increasingly mediate clinical training, immersive simulations become formative spaces in which learners encounter emotionally demanding scenarios. This raises important ethical questions concerning how digitally mediated environments shape stress regulation, agency, and the development of professional dispositions. This exploratory study investigates nursing students' physiological stress responses and interactional behaviors during three trauma-related learning tasks within a virtual reality (VR) emergency care simulation: wound care, auscultation, and chest tube insertion. Six nursing students participated in individual VR sessions, during which electrocardiogram data were collected to assess heart rate variability (HRV) as an index of stress regulation. Synchronized video recordings captured participants' verbal and embodied conduct, enabling a multimodal analysis of coping strategies and task engagement. Results indicate distinct temporal stress patterns across tasks. The wound care task showed the greatest inter-individual variability in HRV, reflecting differences in familiarity and coping strategies. Auscultation elicited moderate physiological engagement, while chest tube insertion was associated with consistently lower HRV, indicating higher task demand accompanied by adaptive regulation rather than excessive distress. Video analysis revealed visible confusion and efforts to regain composure, particularly during the initial wound care task, aligning with physiological findings. These findings suggest that immersive XR simulations function not merely as technical training tools but as environments in which stress regulation capacities central to professional resilience and prudent clinical judgment are enacted and potentially cultivated. A multimodal approach offers insight into how AI-mediated and digitally structured learning environments can be designed to support both competence and psychological well-being. By foregrounding the ethical significance of stress regulation in simulated trauma care, this study contributes to broader discussions on virtue formation, responsible technological design, and digital well-being in contemporary healthcare education.

Aligning large language models with human moral judgment: insights from a subset of the ETHICS dataset

Authors: Alina L. Machidon; Mateja Centa Strahovnik; Vojko Strahovnik; Jonas Miklavčič; Octavian M. Machidon

Institutional Affiliation: University of Ljubljana, Faculty of Computer Science and Faculty of Theology

Abstract:

The alignment of large language models (LLMs) with human moral and ethical values is a critical concern as these models are increasingly deployed in decision-making, content moderation, and advisory roles. Misalignment can lead to outcomes that conflict with societal norms or ethical expectations, emphasizing the need for systematic evaluation and refinement. This study investigates the depth of this alignment using a subset of the ETHICS dataset, comprising 500 instructions across five domains. Using a rigorous protocol, we compared the annotations of a five-person human pool against three state-of-the-art models: ChatGPT 5.2, Gemini 2.5 Pro, and Claude 4.5 Sonnet. Using a moral ontology of ten principles, we identified high surface-level agreement (≈ 0.83–0.94) but discovered disagreement clusters that reveal distinct normative logic between humans and machines. When compared with the original ETHICS dataset, human annotators achieved higher agreement than the LLMs, indicating a more robust alignment with established normative labels. While convergence remained strong in structured domains like Deontology and Utilitarianism, it weakened considerably in Justice and Virtue, a result consistent with the inherent normative complexity of those fields. Our analysis of these disagreements reveals several core patterns, starting with the Virtue domain, where humans consistently identified character traits within social behaviors. In contrast, LLMs frequently denied the presence of a virtue or vice in these same scenarios, suggesting that while humans apply a deeper moral descriptions to everyday actions, LLMs maintain a shallow factual neutrality that fails to recognize the character-building dimensions of social conduct. This divergence extended into the Justice domain, where a distinct legalism bias emerged. LLMs overwhelmingly justified claims based on formal status or institutional licenses while human participants often rejected these institutional justifications in favor of relational reciprocity and emotional desert, validating claims based on the quality of a relationship or a justifiable reaction to unfair treatment. This suggests that the AI's sense of justice is currently anchored in bureaucracy rather than the human experience of fairness. Finally, in Utilitarian trade-offs, the models demonstrated a reliance on pro-social or healthy keyword optimization, such as choosing apricots over ice cream regardless of context. Human annotators, however, prioritized subjective experience and narrative impact, viewing material loss as a less significant negative utility than physical discomfort.

These findings suggest that current alignment strategies produce models that are legalistically consistent but relationally hollow. This gap poses significant challenges for digital well-being, as AI systems may fail to validate the very relational and character-based values that define human flourishing. We conclude by arguing for a necessary shift from rule-based alignment toward a "virtue-centric" approach that accounts for the intentional, emotional, and lived textures of human life.

The impact of online echo chambers in epistemic virtues and vices key for political debate: how online interactions are shaping and shaped by our epistemic character and beliefs about democracy

Author: Francisco Miguel Macías Pozo

Institutional Affiliation: Spanish Ministry of Universities (FPU 23 Contract)

Keywords: Echo Chambers; Virtue Epistemology; Political Epistemology; Political Debate

Abstract:

In digital contexts, our inquiry-related activities that form beliefs can work differently than in offline context. Social media algorithms are designed to exploit or take advantage of human biases and psychological mechanisms. The characteristics of these online platforms can be understood as external circumstances that modify not only or directly our beliefs, but our behavior, attitudes and second-order beliefs, adding characteristics to the first such as greater confidence or emotional intensity or prominence in the subject's identity. Nowadays, the use of digital platforms to learn and share information, opinions and arguments about politics can be even more usual that doing it offline. Politicians, citizens and people from other countries are interacting with each other, debating, attacking or trying to manipulate and produce some beliefs, attitudes or behavior about democracies with people with the same or different beliefs. As will be argued in this article, all of the above can affect positively or negatively with a relevant subset of the internal circumstantial variables that affect belief formation processes: the systems of attitudes or habits that we call epistemic character traits. These can be defined as those characteristics that affect the transmission, maintenance, or generation of knowledge, understanding or other epistemic goods. That is, attributes of a person's character as an epistemic subject that make us excellent or poor learners, researchers, and communicators. This article aims to show the epistemic, social and power dynamics that affect the formation and revision of beliefs of a person who is a member of an echo chamber, as well as their consequences for that person depending on some of their character traits, and ultimately for political debate in liberal parliamentary democracies. Thus, using a methodological approach of political epistemology and the epistemology of virtues and vices, the processes and effects will be analyzed from three pairs of a virtue and a vice that are key for political debates: epistemic humility and arrogance (Tanesini 2021), deference and autonomy (Matheson 2024), and open-mindedness and close-mindedness (Turner 2025). Meanwhile, ethical reflections about how to tackle the challenges for debate in parliamentary democracies will be displayed, either from an individual or institutional perspective. A first research question will be: how do social media characteristics, specially in echo chambers, are affecting the development of the selected epistemic character traits? But the main research question will be: How these epistemic character traits are modifying belief, attitudes and behavior towards political debate, either in online or offline contexts? To answer this, I will first explain the different characteristics of belief formation in relation to political debate in non-digital versus digital contexts, drawing on interdisciplinary scientific literature and even empirical studies, focusing on echo chambers (Nguyen 2020). Secondly, the epistemic character traits chosen for analysis will be defined. The third step will be to argue how these traits can appear or be modified from the inquiry-related activities on social networks with certain attributes. During all the earlier sections, the key points will be applied, developed and connected around to analyze and argue about the consequences on democratic beliefs, attitudes and behavior, and the fourth section will condense and make clear the main argument of the article.

Living in the Cut: Virtuality as a Post-Cinematic Environment of Moral Formation

Author: Jonas Miklavčič

Institutional Affiliation: University of Ljubljana, Faculty of Theology

Keywords: virtuality; post-cinema; virtue ethics; self-presentation; digital culture

Abstract:

This paper argues that contemporary virtuality is not merely digitally mediated but post-cinematically organized. Social media, videoconferencing platforms, and algorithmically curated feeds disseminate across everyday life operations once associated with classical cinema: framing, mise-en-scène, editing, and the direction of the gaze. The virtual is philosophically significant not chiefly because it is less real than the "real," but because it increasingly gives life the form of a scene: experience is anticipated as something visible, shareable, and aesthetically coherent. The central claim is that digital platforms function as post-cinematic environments of moral formation, as they prearrange the conditions of visibility, recognition, self-presentation, and response. The user is shaped as a character, others appear as audiences, and the algorithm assumes the role of a dispersed direction of appearance. The cut has left the editing room and entered life. The paper asks how such an environment reshapes virtue, public space, and everyday habits of self-monitoring.

Trustworthy AI and the King Midas Problem

Author: Thomas Mitchell

Institutional Affiliation: Institute for Ethics in AI, Corpus Christi College, University of Oxford, UK

Abstract:

Should we make AI that is trustworthy, or merely reliable? Philosophers standardly distinguish between the two. Reliability is performing well in a task or role, whereas trustworthiness requires something more: encapsulated interest, responsiveness to dependency, or fulfilling commitments, for example. Traditionally, only persons are thought capable of trustworthiness; tools and machines, including AI, can only be reliable. For current purposes, I assume that AI can be trustworthy on at least some plausible accounts and focus on the desirability thereof. In creating ever more advanced AI, is there any good reason for designing AI that is trustworthy rather than reliable? I propose and defend a positive answer. Although for many specific uses, merely reliable AI may suffice, trustworthiness is necessary for resolving the King Midas Problem, an aspect of the broader value alignment problem. The Problem, named for the mythological figure who wished for all he touched to turn to gold, only to regret it upon realising that he could no longer eat or drink anything, is the possibility that a sufficiently advanced AI may be able to give us exactly what we ask for, but we realise too late that it has undesirable consequences. Bostrom's paperclip maximiser is a particularly extreme example. King Midas situations can be defined using the following conditions: 1. Unforeseeability: An instruction to the AI will lead to consequences that are unforeseen at the time 2. Harmfulness: Some of those unforeseen consequences are severely harmful 3. Unpreventability: By the time the harmful consequences are realised, it is too late to prevent them 4. Permanence: The harmful consequences are irreversible once they have occurred A natural reaction is to ensure that at least one of the conditions remains unfulfilled. For instance, we could pour resources into predicting exactly how an AI would react before giving it an instruction, or design the AI to work very slowly, so that we would have time to stop it upon realising that there will be harmful consequences. But such avoidance strategies would severely mitigate the usefulness of the AI. An actual solution would involve an AI that perceives and is guided by our interests, not just carries out our instructions. It would act on our behalf, while using its own judgement about how best to do so. This would require exercising discretion in interpreting instructions, responding to the ways in which we depend on it, and occasionally resisting instructions that are not in our best interests. These are all features of trustworthiness. I shall defend two theses: that trustworthy AI is sufficient for resolving the King Midas Problem; and that it is also necessary for resolving it. These will be strong considerations in favour of the desirability of trustworthy AI.

Am I the One Who Decides? Determinism in Profiling and Recommender Systems in the Age of AI

Author: Mátyás Nagy

Institutional Affiliation: Ecumene Doctoral School at Babeș-Bolyai University in Cluj-Napoca

Abstract:

In the age of artificial intelligence, the fundamental conditions of individual decision-making have changed. Originating in the data-driven economy and digital marketing, profiling has far exceeded these initial domains of application and poses significant challenges to the possibility of autonomous decision-making. From the perspective of users in the digital environment, the ethical question of decision-making has become pressing: “Am I the one who decides?”

My research addresses precisely this critical ethical issue: how algorithmic profiling and recommender systems shape autonomous decision-making and its underlying conditions in the age of AI. In the process of profiling, algorithms do not merely describe or represent individuals on the basis of their digital footprints. Rather, they go beyond this descriptive function. Recommender systems and filter bubbles tailored to personal profiles may narrow the range of options necessary for autonomous choice or actively influence and steer users in particular directions.

Accordingly, the central questions of my ethical inquiry are the following: (1) How do the concepts of profiling and recommender systems relate to one another? In what ways does artificial intelligence shape these mechanisms from an ethical perspective? (2) How do profile-based recommender systems transform users’ autonomous decision-making, and what possibilities exist for ethical orientation in this context? (3) To what extent can we speak of the challenge of determinism in relation to profiling, recommender systems, and human decision-making, and in what context does this problem arise?

In my analysis, I examine these phenomena of profiling and recommender systems using the analytical frameworks of both legal and technology ethics. Within this ethical analysis, I also seek to identify specifically AI-related factors within profiling.

Does algorithmic profiling merely influence individual choices, or does it structurally reshape the very conditions of decision-making? How, then, should we answer the question: “Am I the one who decides?”

Virtues in joint reasoning with LLMs

Author: Jakob Ohlhorst
Institutional Affiliation: Applied Ethics – RWTH University Aachen

Abstract:

Interacting with LLMs as a deliberative practice both for practical and theoretical questions has grown in significance. People ask LLMs for advice about what to do in particular situations, how to do it and so on in practical concerns. On the theoretical side, LLMs are asked for facts and figures, for arguments for and against positions, explanations of difficult issues. Significantly, the implementation of LLMs in a chatbot format as well as their linguistic competence leads to them being treated like human agents with theoretical, practical, and moral judgment. This talk takes a virtue theoretical perspective on this phenomenon. For a good outcome that is not just lucky, virtues both moral and epistemic need to be manifested. However, whose virtues? Is the only potentially virtuous agent here the LLM's user? If so, can their behavior *[or behaviour, if using UK English]* be virtuous if they treat the LLM like a knowledgeable agent, as we almost cannot avoid given the systems' design features?

I will instead consider further potential loci of virtue in such cases: First, given the behavior of many users, could an LLM be meaningfully called virtuous in any sense? Consequentialist virtue theories might offer some promise in this direction. Second, beyond the human user alone, might the user-LLM dyad form its own locus of extended virtue? This would be in analogy to collective and extended virtues, where the LLM serves as a kind of cognitive scaffolding.

Given these three loci, we receive four options to explain successful LLM-use to evaluate:

- A) there can be no virtue in human-LLM interaction in either the human or the LLM – successful LLM use is not possible;
- B) the human agent must be virtuous in such interactions;
- C) the LLM can and needs to be virtuous in such interactions;
- D) successful LLM use can only be understood in terms of the virtues of the extended dyad of human user and LLM.

I will argue that option D) is the most fruitful in theoretical terms, as it tracks a host of epistemic practices. However, it also comes with great risks of practical and moral de-skilling, that need to be carefully investigated. As a result, it may turn out that option B) is virtue-ethically more preferable.

Virtual Acknowledgement: A Cavellian Approach to Empathy and Immersive Technologies

Author: Pietro Phelan

Institutional Affiliation: Università degli Studi di Modena e Reggio Emilia (Italy)

Abstract:

In recent years, the notion of 'virtual empathy' has frequently been employed to describe the promise of immersive technologies, particularly VR, which are often said to intensify affective identification beyond the limits of traditional media. Although these accounts are articulated in different ways, a significant strand in the discourse on immersive media associates their ethical and socio-political potential with the reduction of distance and the enhancement of presence.

Recent philosophical discussions have questioned whether empathy provides an adequate framework for understanding our relation to others in digital environments. The aim of my research talk is to propose an original approach to these problems, drawing on Stanley Cavell's philosophy and, more specifically, on his concept of 'acknowledgement'. In his essay Notes Mostly About Empathy, Cavell argues that the ethical encounter with others does not consist in affective fusion, but in accepting the irreducible distance that structures our existence. Whereas empathy seeks to bridge the gap with the other, acknowledgement requires learning separateness. In this respect, the arts - particularly theatre and cinema - function as privileged sites of acknowledgement education: in works such as The Avoidance of Love and The World Viewed, Cavell claims that fictional characters become others whom we genuinely confront only insofar as we accept the spatial and temporal separation that prevents our full presence to them and their suffering.

By bringing this philosophical framework into dialogue with contemporary artistic debate and selected immersive practices, my talk proposes 'virtual acknowledgement' as a heuristic category for reassessing the ethical aspirations attached to immersive technologies. Rather than equating ethical efficacy with intensified presence, such a category allows us to examine how immersive works negotiate the tension between proximity and distance. This perspective allows us to reinterpret works such as Carne y Arena, which, while immersive, foreground the interplay between virtual presence and physical absence.

On this basis, two main claims can be articulated: (a) the discourse and enactment of virtual empathy risk producing a miseducation in acknowledgement, potentially constraining the ethical and socio-political aims attributed to immersive art; (b) Cavell's philosophy suggests that presence and immersivity must be balanced by an explicit thematization of distance and separateness if immersive technologies are to sustain a genuinely relational ethical framework.

Cultivating Intellectual Character in AI-Mediated Classrooms

Authors: Madeleine Potoskie

Institutional Affiliation: Department of History and Philosophy of Science, University of Pittsburgh

Abstract:

The rapid integration of generative AI tools into educational settings has prompted widespread concern about plagiarism, authorship, and academic dishonesty. Institutional responses have largely focused on rule enforcement and detection technologies. This paper argues that such approaches misconstrue the central ethical issue. The deeper pedagogical question is not merely whether students misuse AI, but how AI-mediated practices shape the formation of intellectual character. I propose reconceiving academic integrity as a matter of cultivating academic values beyond mere honesty, particularly curiosity, intellectual responsibility, and reflective judgment, rather than simply preventing rule violations. While generative AI can function as a shortcut that bypasses intellectual struggle, it can also create a distinctive dialogical space in which students must formulate questions, refine prompts, evaluate outputs, and make deliberate choices about what to accept or reject. Properly structured, this interaction can foster meta-cognitive awareness and disciplined curiosity. Drawing on my own classroom practices that require transparency and reflective disclosure of AI use, I argue that engagement with generative AI need not undermine intellectual virtue. Instead, when embedded within pedagogical norms that emphasize responsibility and self-awareness, AI can become a site for cultivating the very academic values it is often thought to threaten. The classroom thus becomes a testing ground for broader questions about epistemic agency in digitally-mediated environments.

Epistemic Courage, Trust, and Faith: In AI, and Because of AI

Author: Dušan Rebolj

Institutional Affiliation: University of Ljubljana, Faculty of Arts

Abstract:

The contribution adds to my previous work on epistemic courage as a virtue involved in thinking about, and addressing the issues related to the development and adoption of AI technologies. In this context, I will explore the link between epistemic courage, trust, and faith. The argument will relate to two wider topics: firstly, the debate on making AI trustworthy or reliable (as summarised by Bareis, 2024); secondly, on the importance of inter-human trust given the way in which careless or malicious use of AI endangers epistemic backstops, e.g. the increasing unreliability of unedited video due to the proliferation of deepfakes (Risse, 2023). I will test the following claims: the concept of warranted trust entails epistemic courage, defined as the persistence in upholding one's evidential commitments despite difficulties or risks to one's mental state; under certain conditions, epistemic courage plays this role analogously in the maintenance of warranted trust in AI, and in other people in the wake of AI's pernicious epistemic effects. The first claim assumes the following definition of warranted trust. For trust to be warranted, it must satisfy three necessary and jointly sufficient conditions. It must be plausible (conditions must obtain in which the requisite mental attitude can and does develop), justified (rational regarding the knowledge of the trustee's trustworthiness, and the value trust is supposed to realize), and well-grounded (directed at a trustworthy agent, but perhaps not successfully) (McLeod, 2023). I will demonstrate that in many domains, not just the human relationships with AI and those arising because of AI's effects, epistemic courage can be involved in sustaining all three aspects of warranted trust. In testing the second claim, I will pay particular attention to the issue of warranted trust's good grounding. Namely, the trustworthiness of an agent is predicated, to some degree, on the transparency of the agent's mental processes. To the extent that these are inscrutable, and trustworthiness cannot be assessed, the act of relying on the agent ceases to be one of trust and becomes one of faith (as suggested by Tillich, 2000). Because choosing to rely on an inscrutable agent involves difficult reckoning with the (bases of) one's evidential commitments, epistemic courage is necessarily involved – both in dealing with AI, and in dealing with people who may be deceiving or misguiding us by means of AI.

The Problems of AI, The Reconfiguration of Education, and Intellectual Character Paradigm of Education

Author: Carl Sohmer
Institutional Affiliation: UC Irvine
Keywords:

Abstract:

The rapid development of artificial intelligence has precipitated a crisis in higher education, fundamentally challenging traditional conceptions of the nature and aim of university education. As AI systems demonstrate increasing capacity to perform skilled tasks and generate expertise more efficiently than human graduates, educational paradigms centered on skill acquisition and knowledge transmission appear increasingly untenable. This paper examines how the widespread adoption of AI by students—with recent studies indicating that 90% use AI academically—has not merely created problems of academic dishonesty but has exposed deeper philosophical questions about educational purpose and institutional mission. I argue that the prevailing models of higher education are inadequate to address this crisis because they lack a coherent philosophical foundation for distinguishing what AI can replicate from what genuinely constitutes educational value. In response, I propose the Intellectual Character Paradigm (ICP) as a philosophically grounded alternative that centers education on the cultivation of intellectual virtues and the mitigation of intellectual vices. ICP offers several advantages: it articulates a vision of education that AI cannot compromise; it provides principled criteria for determining appropriate AI integration across disciplines; it presents an aspirational account of human flourishing that is non- utilitarian yet attentive to vocational preparation; and it offers a paradigm for the proper use of AI. I will then conclude that while not a panacea for all challenges posed by AI, ICP represents a more coherent and defensible paradigm than the status quo, particularly when universities must articulate explicit commitments about the fundamental nature and aims of education. Roadmap: Section I examines how AI undermines traditional educational paradigms based on skill acquisition and expertise. Section II analyzes the practical challenges of academic integrity and the widespread adoption of AI tools among students. Section III considers institutional pressures to integrate AI and the demand for AI fluency. Section IV presents the Intellectual Character Paradigm as a philosophical alternative, exploring its theoretical foundations and practical implications for higher education. Section V addresses potential objections and limitations, then concludes with implications for educational policy and practice.

Reasons and Ethics Training of LLMs

Authors: Vojko Strahovnik; Mateja Centa Strahovnik

Institutional Affiliations:

Vojko Strahovnik: University of Ljubljana, Faculty of Arts and Faculty of Theology

Mateja Centa Strahovnik: University of Ljubljana, Faculty of Theology

Keywords: AI alignment; LLM ethics training; normative justification; reasons-reponsivness; sensitivity to context

Abstract:

Current approaches to AI alignment and Large Language Model (LLM) ethics training fail to meet the philosophical requirements of genuine normative justification, substituting statistical preference satisfaction for the kind of reasons-tracking that accountability usually demands. This paper argues that reasons-responsiveness — the capacity of an agent or system to recognize, process, and respond to the normative force of reasons — constitutes the foundational criterion for ethically adequate AI alignment. Prevailing training paradigms, being largely outcome-focused and English-centric, systematically foreclose reasons-responsiveness by homogenizing ethical diversity and suppressing the contextual sensitivity through which reasons acquire their normative weight. Drawing on the philosophical tradition of reasons-responsiveness, we propose a shift from statistical alignment to a reasons-based framework in which LLMs are trained to trace their outputs to explicit, culturally grounded, and ethically justified rationales. On this account, moral competence is not a primitive target of training but an emergent property of systems that are genuinely responsive to reasons across diverse linguistic and cultural contexts. This talk also presents an ongoing project to develop a reasons-based ethics training framework for GaMS, Slovenia's open-source large language model. The project works by adapting an existing ethics dataset — the ETHICS dataset — through human annotators. These annotators enrich ethical scenarios by adding explicit moral rationales drawn from a range of normative principles, e.g., harm prevention, honesty, promise-keeping, justice, and respect.

Moral AI and Moral Upskilling – The Procedural Artificial Moral Assistant

Author: Marco Tassella

Institutional Affiliation: Institute of Philosophy (IFZG) – Zagreb, Croatia

Abstract:

Recent debates in digital ethics have revisited the question of whether AI systems can meaningfully assist humans in their moral endeavors without undermining their autonomy. This paper proposes a procedural turn: as Artificial Moral Systems (AMS) assume a growing role in decision-making, the focus should shift – from answer-giving moral advisors, to process-centered companions able to scaffold human deliberation. Instead of delegating decisions to machines, these moral companions could help human agents sustain reflective judgment under uncertainty, while preserving choice authorship, accountability, and ethical pluralism. Building on proceduralism and moral expertise, this proposal envisions a new form of AI-mediated moral education, by introducing a Procedural Artificial Moral Assistant (PAMA) aimed at enhancing human ethical judgment by upskilling individual moral capacities. This procedural-epistemic stance situates PAMA against the excesses of the three most influential paradigms: AI-substitution (precision at the cost of moral deskilling), AI-advice (prone to conservatism and paternalism), and AI-interlocutor (which risks over-intellectualizing moral deliberation). Drawing on recent proposals, this presentation will underscore the importance of "moral mentoring" through critical reflection, value coherence, and narrative engagement to cultivate individual moral progress with durable motivation. With the system's support, users could develop their abilities to notice and contrast biases, conceptual gaps, and value tensions while retaining control over the outcomes of their decisions. At the same time, PAMA integrates insights from moral psychology and moral education: a robust focus on rational discourse is complemented by imaginatively rich encounters to strengthen the motivational dimension of moral growth and action. Two design commitments follow. First, non-substitutive support and theory-agnostic rigor: extending on the procedural moral enhancement model, user assistance should remain method-transparent and compatible with multiple normative frameworks, while enabling users to test and refine their value hierarchies without collapsing disagreement into moral skepticism or ethical inflexibility. Second, gradual and extended learning: pre-/post-decision review, counterfactual exploration, and curated exposure to perspective-taking narratives should operate as complementary levers for procedural upskilling and character formation, while guarding against technological overreliance. In addressing widely discussed concerns, such as bias and transparency, context-sensitivity, emotional manipulation, and AI-overreliance, the PAMA model proposes guardrails such as a methodology for explicable deliberation, a plural corpus of both ethical theories and narrative materials, and an interaction model that emphasizes user authorship and discourages AI deference. Rather than forecasting a "moral machine", this account recasts moral AI as a companion in human ethical reasoning, by integrating procedural enhancement with morally formative education. The contribution aims to be a theoretically disciplined yet practically oriented framework for AI-supported moral growth that treats agency, pluralism, and motivation as co-equal design constraints, inviting multidisciplinary and rigorous dialogue across philosophy, moral psychology, and the ethics of human-AI interaction.

Aesthetic Injustice without Participants?

Author: Borut Trpin

Institutional Affiliation: University of Ljubljana and University of Maribor

Abstract:

Recent work on aesthetic injustice argues that aesthetic judgment can wrong others by marginalizing or silencing them as aesthetic agents. In digital culture, concerns about AI-generated imagery are frequently paired with character-level prescriptions, such as calls for aesthetic humility, deference, or "listening" in response to suspected harms. This paper offers a scope result that clarifies when such virtue-inflected responses are apt. I argue that leading accounts of aesthetic injustice (e.g., Nanay, Lopes, Fraser) rely on a largely implicit condition: aesthetic injustice presupposes aesthetic participants whose aesthetic capacities ground normative standing, i.e. authority over how an object is to be interpreted or evaluated within a practice, standing that can be overridden, denied, or displaced. AI-generated imagery provides a limiting case. It is now ubiquitous and routinely subject to aesthetic appraisal (beautiful/banal, offensive/compelling), yet many paradigmatic uses (high-volume, attribution-light generation for feeds, thumbnails, ad variants, placeholders) lack any agent who plausibly occupies the role of an aesthetic participant. Against a content-first view on which representational harm alone suffices for aesthetic injustice, I argue: (i) AI systems are not aesthetic participants, since they do not occupy a standpoint within a practice that can bear or claim aesthetic authority; (ii) contributors to training data, while causally implicated, do not thereby retain standing over the interpretation or evaluation of particular outputs; and (iii) online "prompting" communities typically institute norms of procedural competence rather than practice-relative authority over aesthetic meaning. Hybrid cases in which AI functions as a tool within a recognisably human artistic practice mark boundary conditions rather than counterexamples: where agents claim and exercise aesthetic authority within shared evaluative practices, aesthetic injustice becomes possible again. The upshot is not that AI-generated imagery is ethically benign. Rather, many of its central wrongs, such as bias reproduction, stereotyping, and extractive appropriation, are better diagnosed as structural, institutional, or economic injustices. Misclassifying them as aesthetic injustice risks misdirecting normative attention: it encourages interpersonal virtue-responses (deference, humility) where what is primarily needed is institutional accountability and responsible governance (documentation, auditability, explainability, and escalation). This has implications for how we theorize virtue in digital culture: virtues like humility and deference presuppose relational practices among agents with standing, whereas many AI-mediated aesthetic contexts lack such relational structures. In these domains, ethical formation must be pursued at the level of institutional design rather than interpersonal virtue cultivation.

Building GaMS: Addressing Data Scarcity and Cultural Alignment in a Less-Resourced Language

Author: Domen Vreš

Institutional Affiliation: University of Ljubljana, Faculty of Computer and Information Science

Abstract:

While Large Language Models (LLMs) have become the primary interface for digital interaction, they are predominantly trained on Anglocentric datasets. In this talk, I present GaMS (Generative Model for Slovene), an open-source LLM specifically developed for the Slovene language. I will analyze the challenges inherent in working with a less-resourced language. One important challenge is the lack of authentic Slovene datasets for post-training, leading to the adaptation of existing Anglocentric datasets. I will present our methodology for cultural adaptation, moving beyond simple translation to ensure the model respects local cultural nuances. Finally, I will explain how safety and ethical alignment are integrated into the LLM training process and demonstrate why this remains an open challenge for the Slovene linguistic context.

Program Committee:

Vojko Strahovnik

Department of Philosophy, Faculty of Arts, University of Ljubljana

Head of the Centre for Human-Centred Artificial Intelligence and the Ethics of New Technologies

Mateja Centa Strahovnik

Faculty of Theology, University of Ljubljana

Leader of the research programme The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights

Diana Daly

College of Information Science, University of Arizona

Associate Dean, Graduate Academic Affairs and Student Success

Ivan Cerovac

Department of Philosophy, Faculty of Humanities and Social Sciences, University of Rijeka

Leader of the research project Epistemic Democracy in a Digital Era, HRZZ

Organized by:

Centre for Human-Centred Artificial Intelligence and the Ethics of New Technologies
Faculty of Theology, University of Ljubljana

2026, Faculty of Theology, University of Ljubljana