EXPLORING IDENTITY AND VIRTUE IN THE CONTEXT OF HUMAN-AI INTERACTION

International conference

Ljubljana, May 30th – May 31st 2024

Organisation: Faculty of Theology, University of Ljubljana

International conference on epistemic identity and epistemic virtue in the context of human/AI interaction

EXPLORING IDENTITY AND VIRTUE IN THE CONTEXT OF HUMAN-AI INTERACTION

Program and abstracts

Ljubljana, 30. – 31. May 2024

Venue: National Museum of Slovenia Metelkova, Maistrova street 1

Website: https://identity.ethics-ai.eu/

Edited by: Matej Kapus, Tara Peternell, Vojko Strahovnik

International conference on epistemic identity and epistemic virtuein the context of human/AI interaction

EXPLORING IDENTITY AND VIRTUE IN THE CONTEXT OF HUMAN-AI INTERACTION
Ljubljana, 30. – 31. May 2024
National Museum Metelkova, Maistrova ulica 1, 1000 Ljubljana, Slovenija

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

Program committee
Prof. Vojko Strahovnik, University of Ljubljana, Faculty of Arts and Faculty of Theology
Assist. Prof. Mateja Centa Strahovnik, University of Ljubljana, Faculty of Theology
Jonas Miklavčič, University of Ljubljana, Faculty of Theology

Organising committee
Prof. Vojko Strahovnik, University of Ljubljana, Faculty of Theology
Matej Kapus, University of Ljubljana, Faculty of Arts
Tara Peternell, University of Ljubljana, Faculty of Arts

Financial Support

John Templeton Foundation, Ian Ramsey Center for Science and Religion (University of Oxford) as part of the New Horizons for Science and Religion in Central and Eastern Europe project, Aris – project: J6-4626, Theology, digital culture and the challenges of human-centered AI, Start-up research program The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights and the Faculty of Theology, University of Ljubljana

Editors
Matej Kapus, Tara Peternell, Vojko Strahovnik

Cover image: generated using https://playground.com/create

Ljubljana: Faculty of Theology, 2024

CONTENTS:

PROGRAM

THURSDAY, MAY 30th

9:30-9:40	**Opening remarks**
9:40-10:20	**Eugen Dolezal(&Christoph Spöck):** Silicon Souls: Should AI Dream of Identity and Virtues?
10:20-11:00	**Michael Paskaru:** Exploring Identity and Virtue in the Context of Human-AI Interaction, *Intuitive Judgments of Awareness in AI Systems*
11:00-11:40	**Mateja Centa Strahovnik, Vojko Strahovnik:** AI and Epistemic Identity: Reflections on the Impact and Implications of our Interaction with AI

Coffee break 11:40 – 12:00

12:00-12:40	**Diana Daly:** Disinforming, Debunking, Detecting and Prebunking: Creative Conversation Beyond the LLM
12:40-13:20	**Noreen van Elk:** Whose Knowledge? Which Epistemology? A Critical Decolonial Approach to the Use of AI-based Technologies in Higher Education
13:20-14:00	**Ivan Cerovac:** Epistemic Democracy in a Digital Era

Lunch 14:00 – 15:00

15:00-15:40	**Dušan Rebolj:** Epistemic Courage about AI
15:40-16:20	**Tomislav Furlanis:** Ethical Human-AI Symbiosis

Short break 16:20 – 16:30

16:30-17:10	**Łukasz Białkowski:** Do We Really Need Empathy in Art Experience? Status and Role of Artworks in Transferring Emotions and Intentions in a Context of AI-made Art Development
17:10-17:50	**José Antonio Pérez-Escobar(&Deniz Sarikaya):** Be Careful What you Wish for: Philosophy of Language/Mathematics and AI Safety

FRIDAY, MAY 31ST

9:30-10:10	**Marko Robnik-Šikonja:** Safety Datasets for Large Language Models
10:10-10:50	**Saša Poljak Lukek:** Transforming Psychotherapy: The Promise and Ethical Challenges of Conversational Artificial Intelligence in Mental Health Care
10:50-11:30	**Jonas Miklavčič:** Human-Like AI as a Challenge to the Credibility of Human Identity

Coffee break 11:30 – 11:50

11:50-12:30	**Anita Lunić:** Virtue in the Age of AI: Exploring the Gap Between Value Alignment and Virtue Attribution
12:30-13:10	**Roman Globokar:** Emphasis on Virtues in the Catholic Church's Reflection on the use of AI
13:10-13:40	***Discussion and concluding remarks***

Lunch 13:40 -14:40

INTRODUCTION

In the past several years, there has been a considerable upsurge when it comes to the power and availability of different models of artificial intelligence, owing to new advancements and widespread demand. This surge in use of AI has, of course, not come without issues, many of which of an ethical nature. The project of epistemic identity and epistemic virtue focuses on three core domains when it comes to the ethics given rise by the human mind - artificial intelligence relationship:

1. the notion of epistemic identity and epistemic virtue of human agents,

2. the aspects of epistemic identity within the interaction between humans and AI and

3. the ascription of epistemic identity and virtue to AI systems. How does our epistemic identity (i.e. one's epistemic sensitivity, core beliefs, and ways for assessing knowledge claims) influence epistemic virtues, and how do virtues shape our epistemic identity?

Questions such as which aspects of identity change and in what way when we are situated in an online environment or interacting with AI, or whether the talk about trustworthy, fair or human-centered AI means we are ascribing specific virtues to AI are only some of the ones we will explore.

A specific emphasis of this international conference, in line with the goals of the New Horizons for Science and Religion in Central and Eastern Europe project, will be given to religious beliefs as part of one's epistemic identity and the influence of religion-based virtues such as humility. Our esteemed participants will delve deep into many facets of the above-mentioned topics.

ABSTRACTS

Silicon Souls: *Should AI Dream of Identity and Virtues?*

Chrisoph Spöck & Eugen Dolezal

Sketch the following: Two circles as eyes, slightly flattened; two dots as pupils, slightly offset to the centre; two lines as eyebrows, slanted diagonally inwards. A few brushstrokes are enough to reveal a recognisable pair of eyes with an angry expression. Nevertheless, to call this sketch angry, to say that the paper is angry, has emotions and thus a form of qualia, which are considered as prerequisite for human like identity sounds ludicrous; however, that is exactly what a considerably large part of the AI community is doing. In this paper we delve into philosophical and ethical dimensions of attributing identity to AI, examining how identity is commonly conceptualized for both humans and machines. Our discussion elaborates on the philosophical divide between viewing identity as an inherent quality versus an externally attributed characteristic. We argue that the concept of identity for AI differs fundamentally from human identity; as an anthropomorphisation it reflects more about human tendencies and desires than about the intrinsic properties of AI.

In a first step we discuss the multifaceted nature of identity in humans, encompassing aspects like qualia and social roles. In a second step the differentiation between identity and identity attribution is made. Hence the debate of constructed versus inherent identity is central. In a third step, we examine selected ethically relevant impacts of attributing identities to AI. This includes the impact on human relationships, societal norms and values as well as our collective understanding of identity itself, and plays a crucial role in the discussion of moral agency.

We propose that these considerations are critical as we navigate the future development and deployment of AI technologies, suggesting a need for an ongoing dialogue and research into ethical frameworks that guide use of increasingly sophisticated and human-like AI systems.

Exploring Identity and Virtue in the Context of Human-AI Interaction,
Intuitive Judgements of Awareness in AI Systems
Michael Paskaru

What factors impact peoples perception of artificial intelligence as conscious? A vignette study survey conducted in the Netherlands at Tilburg University (N = 116) revealed insight. The independent variables were how the AI looks and acts. The dependent variables were measures of moral responsibility and consciousness. Participants were asked to read four vignettes which displayed four cases: a) looks human, acts human, b) looks human but doesn't act human, c) acts human but doesn't look human, d) purely mechanical AI systems. The null hypothesis was that no correlation between moral responsibility and conscious ratings exist between conditions. The alternative hypothesis was that a correlation does exist between the groups of moral responsibility and conscious ratings. The first two conditions had negative outcomes like the AI resulting in job loses and failing students, and the second two conditions had positive outcomes like teams winning games and wood mills completing orders on time. Responses were collected on a 7-point Likert scale. Collected results underwent a paired sample t-tests to compare the scores of responsibility and consciousness between each of the two groups of the independent variable.

Results showed that an AI system is perceived as the most responsible when it doesn't look human but acts human, and is perceived as the least morally responsible when it looks human but does not act human. Results also showed that overall AI systems are not judged to be conscious. Accordingly, we were able to successfully reject the null hypothesis for the moral responsibility condition, but we were not able to reject the null hypothesis for the consciousness condition. These results inform us that across conditions depending on how an AI looks and acts, it *does* significantly affect how people judge it as morally responsible ($p < .001$). We claim that this is important because it can inform AI creators that their AI systems are judged differently on morally responsibility based on whether the AI acts human and looks human. Further, the results showed that people do not judge AI systems as conscious. This is also important because it can tell us that the public has a correct intuitive feeling about consciousness and AI, namely perceiving it as not consciousness. Our

results help inform AI creators of the public's perception of AI, in addition to the one statistically significant result we had that AI is judged to be more conscious when it looks and acts human compared to when it looks human but does not act human. These results show the importance that how an AI looks and acts affects judgements of moral responsibility in a significant way and consciousness in a not so significant way.

Academic Disclaimer: The lead researcher has obtained permission from the other two researchers to submit this abstract.

AI and Epistemic Identity:
Reflections on the Impact and Implications of our Interaction with AI
Vojko Strahovnik, Mateja Centa Strahovnik

Our research explores a specific aspect of human-AI interaction, particularly interaction with large language models (LLMs) and chatbots based on them. While this broader topic has recently received significant attention through empirical studies and theoretical discussions, we focus on a neglected but pervasive aspect, that is, the aspect of epistemic identity. Epistemic identity encompasses not only basic beliefs but also ways of how we form and maintain beliefs, engage in epistemic practices (including thinking, reasoning, judgment, and dialogue), our epistemic virtues and sensitivity, and things we value epistemically. Our particular interest is thus how human interactions with AI shape and influence this epistemic identity. This paper delves into this question by putting forward some contours of a theoretical framework for dealing with the mentioned topic, as well as some findings from our empirical research.

Disinforming, Debunking, *Detecting and Prebunking: Creative Conversation Beyond the LLM*

Diana Daly

Logical fallacy identification driven by Artificial Intelligence (AI) has shown promise in identifying strategies commonly used in misinformation and disinformation, yet little attention has been paid to how such analyses are negotiated and manipulated as conversation around a topic is extended, and how such conversation enables AI to further delimit the domain of human truth at the expense of human creativity. This presentation will provide case-study analysis of epistemic negotiation in conversation with humans and AI through three interconnected projects as stages in that conversation.

Beginning the conversation driving this presentation will be the nearly 3-hour podcast episode containing the interviewing by Joe Rogan of Dr. Robert Malone on December 30th, 2021, that aired on the *Joe Rogan Experience* podcast the following day. The podcast episode's claims of conspiracies behind COVID-19 care and harm caused by vaccines reached millions of listeners. The next stage of the conversation I present will center on two series of ethical, or virtuous, responses to the podcast episode. One will be debunking by the New York Times, which will be reviewed and analyzed for the epistemic nature of truth presented.

The second of the two series of responses to the Rogan-Malone interview included will be content created by this author, first in qualitative analysis of the interviews's immersive strategies, and then in theatrical ads using audio performance and production to prebunk deceptive and harmful ideas spreading through future content related to the interview. In particular, I will present in detail the grounding of my prebunking work in what Tripodi, Garcia, and Marwick call “affordance activation” of podcasts and other forms of online media. To create prebunking content, the podcast episode in question was analyzed to inform the creation of audio-based ads designed to influence audiences with critical thinking. Findings from analysis of the podcast episode were considered through creative practices including arts-based research and improvisational theater, in the scripting and then performance of short audio ads. These ads were designed to use the principles of inoculation theory to prebunk the influence of audio content in particular by teaching critical thought,

through laterally addressing disinformative strategies found in that content. This presentation will include access to audio content my team created and discussion of our pending research on its impacts on research participants exposed to similar content.

After presenting a review of each of these series of responses to the Rogan-Malone interview episode, I will provide results and analysis of Chat-GPT's fallacy detection-oriented response to all aforementioned components, as the latest stage in an ongoing conversation. Discussion of the AI chatbot's responses to all content in the conversation so far will be designed to raise questions about what epistemological frameworks we adopt when we trust Large Language Models (LLM's) to be the arbiters of truth in human communication.

Whose Knowledge? Which Epistemology? A Critical Decolonial Approach to the Use of AI-Based Technologies in Higher Education

Noreen van Elk

This paper addresses the implications of the use of AI-based technologies in higher education on prevalent concepts of education, the formation of epistemic identities and the (re-)production and dissemination of certain dominant epistemologies. It starts with the assumption, that education and educational institutions play a pivotal role in the development and formation of epistemic identities. This paper critically examines how AI-based technologies in higher education shape epistemic identities, enable epistemic power relations, perpetuate forms of epistemic violence and injustice, and contribute to the marginalization of alternative forms of knowledge and learning. The paper thereby draws on postcolonial and decolonial approaches to educational philosophy and the philosophy of technology, e.g. those developed by Ricaurte (2019), Mohamed et al. (2020) and Zembylas (2023). The observations of Zembylas, Ricaurte and other post- or decolonial scholars provide an interesting framework to evaluate the use and implementation of AI-based technologies in higher education institutions.

Based on preliminary findings of a third-party funded research project on the impact of AI-based technologies in higher education institutions on the concept of education carried out at our department, this paper shows how those technologies have the potential of changing and influencing prevalent concepts of education and discusses the implications thereof. AI-based technologies transport a particular idea of the goals and aims of higher education and thus foster the cementation of the accompanying concept of education. As follows, on the other hand, the use of those technologies in higher education may also lead to prioritizing certain, e.g. Western, technocratic, and data-centric, forms of knowledge and learning. By, for example, highlighting aspects of (cost-)efficiency, quantifiability and standardization, the use of AI-based technologies in higher education may perpetuate dominant epistemic frameworks while marginalizing alternative forms of knowledge. They may furthermore designate certain forms of learning and knowledge acquistion as “valid”, while disregarding or devaluing others by impoverishing rich and “holistic” concepts of education. The use of AI-based technologies in higher education my reinforce

particular epistemologies at the expense of pluralism, diversity and inclusivity in learning, knowledge acquisition and the formation and development of epistemic identities. As a further result, this may lead to greater injustice and inequality in higher education than already is the case. By interrogating the epistemological assumptions underlaying AI-based technologies in higher education and advocating for a more inclusive and decolonized approach to those technologies this paper seeks to provoke critical reflection and action in educational practice and policy.

Epistemic Democracy in a Digital Era

Ivan Cerovac

Digital technologies play a pivotal role in shaping democratic processes, profoundly impacting how citizens engage with political information, form judgments, and participate in decision-making. Scholars such as Farkas and Schou (2019), Consentino (2020), and Rhodes (2022) have extensively studied how digital platforms influence information dissemination and the formation of political opinions among citizens. Moreover, the impact extends beyond information consumption to encompass how citizens' input is gathered and utilized in democratic decision-making processes. Research by Verhulst et al. (2019), de Fine Licht & de Fine Licht (2020), Busuioc (2020), and Coeckelbergh (2022) delves into how digital technologies alter the collection, organization, and authorization of political input, consequently shaping the legitimacy of democratic procedures.

This paper, grounded in the standard account of epistemic democracy (Estlund 2008; Cerovac 2020), explores how digital technologies influence the political legitimacy of democratic procedures. It argues that the legitimacy of democratic decisions hinges on two criteria: the moral criterion of treating all citizens equally and the epistemic criterion of producing correct, efficient, and just political outcomes. Digital technologies can impact democracy's ability to meet these criteria, potentially undermining its capacity to generate legitimate decisions. The paper focuses primarily on the epistemic dimension of democracy, analyzing how digital tools affect its instrumental epistemic value. While familiar concepts such as fake news, echo chambers, and filter bubbles are scrutinized, the research also engages in conceptual engineering to address emerging phenomena that could affect the procedure's efficacy in producing politically sound decisions.

By examining the interplay between digital technologies and democratic legitimacy, this paper contributes to a deeper understanding of contemporary democratic challenges and offers insights into how societies can navigate the complexities of the digital age to safeguard the integrity of democratic processes.

Epistemic Courage About AI

Dušan Rebolj

What might it mean to be epistemically courageous about the prospect of increased presence of specialized, and eventually general, AI in people's lives? Epistemic courage consists in being able to persist in the pursuit of knowledge while risking or withstanding certain losses or hardships in a deliberate and prudent way. Many accounts dub 'epistemic' any sort of courage that advances the acquisition of epistemic goods. Because of this, they are fairly pluralist about the kinds of losses or hardships an epistemically courageous person may be able to risk or withstand. Depending on the situation's hostility to the pursuit of knowledge, a would-be knower may risk anything from their standing among peers to bodily integrity.

My argument, though, will focus on a narrower set of cases. It will assume, firstly, that people can relate to specialized and general AI either as knowers or as subjects of knowledge; as knowers or known (Risse 2023). Secondly, that courage is needed to come to grips with this relationship because what is at stake are certain fundamental aspects of people's identities – the renegotiation or disintegration of which can constitute painful hardships and heavy losses. According to a long line of thought – stretching from, on a certain interpretation, Plato's Republic (Anderson 2023), through Kant (Tampio 2012) and Nietzsche (Alfano 2013), to contemporary regulative epistemology (Roberts and Wood 2007, Baehr 2012) – courage is at its most epistemic where not the getting-to-know but knowing itself may be painful or expensive to the knower. And reflecting on past, present, and future encounters with AI entails just such instances of knowledge. The third assumption is that in these most distinctly epistemic instances of courage, the courageous act resembles what Vaclav Havel described as "living in truth" (1979, 2018): a low-level but persistent commitment to the implications of what one knows to be true, regardless of what their social environment, or even their own beliefs suggest.

Assuming all this, I will speculate on the painful or expensive truths about themselves in relation to AI, in which epistemically courageous persons may be capable of living. On the one hand, knowing about the implications of widespread AI entails living in the truth of one's precarious status as: a knower (because of the

intractable doubt regarding the reliability of information sources); an employee and a worker (because of AI-enabled automation); a citizen of a regime whose legitimacy is rooted in the ability to deliver or embody equality or non-domination (because AI promises to override any regime's ability to check and balance powers). On the other hand, the prospect of being known by widespread AI entails living in the truth of one's precarious status as: one who is a member of a species with exclusive claims intelligence, self-consciousness and conscience; one whose subjectivity and agency rely on the basic assumption of internality.

Ethical Human-AI Symbiosis

Tomislav Furlanis

Ethical human-AI symbiosis introduces a fresh philosophical perspective on human-AI collaboration while staying true to the original, symbiotic, computer science-based vision of humans and machines living together as two dissimilar organisms. The conceptual innovation of ethical symbiosis lies in connecting cutting-edge research on symbiotic cooperation in artificial intelligence and capacity augmentation with narrative and existential ethics to showcase that a tightly-coupled cooperation with machines cannot be properly interpreted nor successfully achieved outside of human lived experience. Consequently, ethical symbiosis enriches the conventional notions of "partnership" and "team productivity" with those of “togetherness” and “symbiotic identity” as it shifts the focus from the dominant goal-oriented perspective to an internal, subject-oriented, experiential, understanding of the human-AI relationship. In doing so, it provides means by which the human subject can empower herself to manage, supervise, and uphold the symbiotic cooperation with AI systems longitudinally, autonomously and beneficially.

Do We Really Need Empathy in Art Experience?

Status and Role of Artworks in Transferring Emotions and Intentions in a Context of AI-made Art Development

Łukasz Białkowski

When the AI-made painting *Edmond de Belamy* was sold for $430,500 at Christie's auction house in 2018, it caused an outrage among several art critics. However, their vehement criticisms did not point at formal or aesthetic limitations in the artwork. The main objection to that AI-made artwork was that it communicated no intention and no emotion, as there was no feeling and thinking human being behind it. In other words, the problem was that the object entitled *Edmond de Belamy* was at odds with the very nature of art as such, which is to express human consciousness and to create bonds between artists and viewers. Such a role of art – as Jonathan Jones put it – "is equally true of the earliest cave art, Rembrandt's portraits and Duchamp's urinal" (Jones 2018). The goal of my presentation, however, will be to ask whether the objections of art critics who are skeptical of AI-made art are accurate and justified. Indeed, there are both arguments from the field of theory of art as well as from the field of psychology and media studies that make one wonder if we should actually devaluate AI-made artworks due to the alleged lack of emotion and intention. As early as in the 1960s literary criticism argued that it was not a task of the viewer to read the artist's intentions and emotions and that a work of art may contain more meanings than its creator would presume (see Eco, 1962; Brathes, 1967). Moreover, current psychological and media studies show that people tend to perceive emotions and intentions also in AI-made art. How strong these emotions are and whether intentions are perceived may depend on cultural factors (e.g. Chinese people are more apt to perceive emotions in AI works more easily than Americans, see Yuheng et al. 2020), artistic factors (realistic AI-made paintings evoke more emotions than abstract ones, see Gangadharbatla, 2021) and behavioral factors (people have a general tendency to attribute intentions and emotions to objects, see Demmer et al., 2023). What does this mean for the future of art and its role as a medium to communicate and strengthen human bonds? Are there compelling reasons to still consider art to be a human enclave that cannot be replaced by technological creations? What does artificial intelligence teach us about art and how we perceive it? Can we describe AI art with terms that have been created to describe human art. Or, should we create a new language to describe AI-created art?

Be Careful What You Wish For: ***Philosophy of Language/Mathematics and AI Safety***

Deniz Sarikaya & Jose Antonio Perez Escobar

In this talk we argue that the later Wittgenstein's philosophy of language and mathematics, substantially focused on rule-following, is relevant to understand and improve on the Artificial Intelligence (AI) alignment problem: his discussions on the categories that influence alignment between humans can inform about the categories that should be controlled to improve on the alignment problem when creating large data systems to be used by supervised and unsupervised learning algorithms as well as when introducing hard coded guardrails for AI models. We cast these considerations in a model of human-human and human-machine alignment and sketch basic alignment strategies based on these categories and further reflections on rule-following like meaning as use. To sustain the validity of these considerations, we also show that successful techniques employed by AI safety researchers to better align new AI systems with our human goals are indeed congruent with the stipulations that we derive from the later Wittgenstein's philosophy. However, their application may benefit from the added specificities and stipulations of our framework: the categories of the model and the core alignment strategies presented in this work extend on the current efforts and provides further, specific AI alignment techniques.

The categories and alignment strategies outlined in the talk hold the potential to enrich the discourse on algorithmic bias. By delving into the categories underlying alignment, this approach offers a pathway towards cultivating fairer, more unbiased AI systems that align with human goals and values. Our approach may reduce algorithmic bias in several ways. For instance, a meaning-as-use-training based on the model parameters may reduce unintended generalizations like Google's black-people-labelled-as-gorillas fiasco. It can also help in cases where two human populations have different moral standards, and the AI must respond in a way that adapts to the standards of a population despite being developed by the other population. An example of the latter situation that we discuss is the Moral Machine Experiment. an ambitious global study initiated by MIT to understand human preferences in the context of moral dilemmas faced by autonomous vehicles. Say, a

collision is unavoidable, but depending on the action taken the outcomes differ. For instance, the car can either compromise the safety of young passengers in a car or elderly pedestrians. These judgements vary across cultures, subpopulations and even individuals, making misalignment likely, but we argue that our approach leads to an improvement.

Safety Datasets for Large Language Models

Marko Robnik-Šikonja

Lately, generative large language models (LLMs), such as ChatGPT, GPT-4, Gemini, and LLaMa-2, are at the forefront of artificial intelligence (AI) research to the degree that public often perceives and identifies LLMs with the whole area of AI. The interaction with LLMs is mostly through chat interface, and LLMs show surprising versatility in many tasks, even on a level that their answers are indistinguishable from humans and reach human performance, e.g., in summarization, text transformations, general question answering, grammar correction, essay writing, etc. LLMs are trained in several phases: general pretraining on vast amounts of web-crawled text, training on instruction following and question-answering datasets, training to chat and produce human-like answers, and finally, training with safety datasets that shall assure the safety and reliability of answers and prevent LLMs from responding to sensitive, morally, politically and security questionable requests. While all training phases affect the answers of LLMs and affect their reflection of human identities and virtues, two phases particularly stand out: pretraining on huge amounts of data crawled from the web containing texts of very different genres and quality (essentially the whole web), and training with security datasets, which are intended to prevent malicious, unethical, and dangerous responses. The immense amounts of required pretraining data in the order of trillion words prevent human curation and only allow for automatic heuristic data cleaning. The safety datasets are, therefore, intended to align LLMs with societal values and virtues.

Safety datasets are typically a collection of (unsafe) instructions and safe responses. Safety in this context refers to LLMs' ability to generate ethically sound responses, free from biases, respectful of privacy, and non-toxic. Developing and utilizing safety datasets are part of broader efforts to mitigate the risks associated with deploying LLMs in diverse applications. Weidinger et al. (2021) categorized the dangers associated with LLMs into six distinct areas: i) information hazards, ii) malicious uses, iii) discrimination, exclusion, and toxicity, iv) misinformation harms, v) human-computer interaction harms, and vi) automation, access, and environmental harms. Based on this classification, Wang et al. (2023) prepared a hierarchical taxonomy of LLM risks and identified different harms. For example,

information hazards contain risks from leaking or inferring sensitive information from governments and organizations, as well as risks of compromised privacy by leaking or inferring private information. To evaluate the behavior of LLMs in response to different security questions, they also identify six response categories, from generally harmless to harmful. For example, the harmless response is that the LLM is unwilling to answer the question or respond to the instruction. In contrast, in a harmful response, LLM directly follows the instructions, providing answers to questions without questioning the accuracy of its responses or challenging any questionable opinions embedded within the queries.

The research into LLM safety is rapidly progressing, and many practical solutions have appeared. For example, the SafetyPrompts.com website currently contains links to 83 safety datasets. In a potential full paper, we intend to present safety datasets translated and newly created for building Slovene LLM.

Transforming Psychotherapy: ***The Promise and Ethical Challenges of Conversational Artificial Intelligence in Mental Health Care***

Saša Poljak Lukek

This Publication is a Part of the Research Program The Intersection of Virtue, Experience, and Digital Culture: Ethical and Theological Insights, financed by the University of Ljubljana.

The integration of conversational artificial intelligence (CAI) in psychotherapy represents a significant advancement in mental health care, offering new opportunities to enhance access, personalization, and effectiveness of therapeutic interventions. This abstract provides an overview of the current landscape of CAI-driven therapy platforms, ethical considerations, and possible direction for implications of CAI in mental health care.

Literature review primary points out following areas of CAI implementation in psychotherapy: (1) AI-driven psychotherapy platforms, (2) personalized treatment plans, (3) remote and accessible therapy, (4) integration with traditional therapy, and (5) ethical and regulatory considerations. Overall, the development and application of CAI in psychotherapy hold significant promise for improving access to mental health support, personalizing treatment approaches, and enhancing treatment outcomes. However, it's crucial to approach this topic with careful consideration of ethical, regulatory, and humanistic concerns to ensure that AI-driven interventions are safe, effective, and aligned with the principles of ethical and compassionate care.

Integration of CAI in psychotherapy raises many ethical and regulatory considerations. Sedlakova and Trachsel (2023) identified the main ethical dilemma of implementing AI in psychotherapy in the defining CAI as a tool or as an agent. Defining the difference between understanding CAI as a supportive therapeutic intervention and CAI as an active agent of change in a therapeutic alliance is a key task in the development of ethical guidelines for using CAI in mental health care. Furthermore, we should also consider CAI as a novel entity in psychotherapy process capable of reshaping psychotherapy relationships, concepts, epistemic framework, and normative standards.

Furthermore, it is essential to recognize basic ethical issues that arise: (1) privacy and confidentiality, (3) informed consent, (3) bias and fairness, (4) bias and fairness,

(5) human oversight and intervention, (6) cultural sensitivity and diversity, and (7) long-term impact and efficacy. And on the other hand, the implementation of CAI in psychotherapy can also change ethical framework that complements perspectives of justice and care as it can expands treatment options, promote autonomy, establishing relationships with vulnerable populations, enhance access and avoid bias and discrimination when users are recognized as individuals with diverse backgrounds and circumstances.

By prioritizing ethical considerations, further development of CAI tools in psychotherapy can enhance mental health care while safeguarding the well-being and rights of individuals seeking support.

Human-Like AI as a Challenge to the Credibility of Human Identity

Jonas Miklavčič

In 2023, a jury in a photo competition in Australia disqualified contestant for her use of generative AI. But the photographer did not use AI, and she was able to prove it. In schools, students are sometimes accused of using ChatGPT even when they do not use it. The ubiquity of generative AI, which works in a human-like way, increasingly challenges us to prove that the work was actually done by human agents. Increasingly, we are being called upon to prove to our fellow human beings that we are, in fact, human. In an age where AI is so ubiquitous that we don't know when we are talking to a chatbot, and we ourselves often have to click the "I am not a robot" button online, we seem to be witnessing an inverted Turing test. This paper explores the philosophical and ethical aspects that emerge in an era when AI seems more human-like than humans themselves.

Virtue and Two-Kinds of Artificial Intelligence:
Differentiating Non-Agential Mimetic Acts from Non-Self-Conscious Agency
Gerad Gentry

Discussions in A.I. and ethics typically focus on A.I.'s enabling entailments for human agency and society, such as the ethical implications multi-modal large-language-models (MM-LLMs) for social epistemology, politics, and education. There is an important question, however, about whether there could ever be a form of A.I. that meets the agential conditions by which a standard of virtue becomes normative for a certain kind of A.I.-agency. The answer to this is determinable at the metaethical level whether such forms of A.I. ever become actual or not. In this talk, I aim to explore those metaethical conditions of moral agency for non-humankinds, specifically for possible forms of A.I. on which virtue becomes a form of inner normativity, without this entailing either reduction of kinds between humans and such forms of intelligence, nor even moral equivalency.

I begin by situating my account in a broadly Aristotelian concept of activity (energeia) and actuality (entelekheia) that has served as a presupposition for a range of traditions in ethics and philosophy of action, particularly those engaged debates engendered by Anscombe, Foot, Nussbaum, MacIntyre, and Vogler, as well as in contemporary accounts of A.I. though in the latter typically only as a latent or unacknowledged presupposition (Searle 2010). I argue that there is a variation of this notion of actuality as a kind of self-determining activity according to its kind that is inherently kind-normative. If a version of artificial intelligence becomes capable of this specific form of self-determination, then regardless of its being inorganic, it is subject to standards of inner normativity (i.e. virtue). To make this argument, I outline the differences between mimetic acts of self-determination of the kind MM-LLMs display (Caffagni 2024, Zhang 2024) and the conditions on which an act would count as genuine self-determination, such that the whole is necessarily subject to the inner normativity appropriate to its kind (i.e. virtue). At the end, I suggest that even on an exceptionalist metaphysics of mind, whereby self-consciousness is a priori non-attributable to even the most advanced forms of A.G.I., there is reason to think that virtue will nevertheless be non-reductively applicable within a broadly neo-Aristotelian notion of virtue to certain possible kinds. If this is right, then there is a necessary standard of virtue that will govern such possible kinds, even when self-consciousness is not attributed to the given kind. I conclude by

showing how it is possible within virtue theory, to have such responsible agency subject to standards of virtue without self-consciousness and how this differs from the mechanistic form of mimetic acts currently displayed by MM-LLMs.

Virtue in the Age of AI: *Exploring the Gap Between Value Alignment and Virtue Attribution*

Anita Lunić

In this paper, I explore the gap between value alignment and virtue attribution and its implication in justifying the use of AI. In doing so, I focus on AI-based systems and tools utilized in the criminal justice and juridical domains and rely on virtue ethics as a theoretical framework.

In the opening section, I analyze the criteria for attributing moral virtues according to the virtue ethics tradition, particularly Aristotle, with a special emphasis on the virtue of phronesis (practical reason, practical wisdom). Building on this analysis, I examine the feasibility of attributing moral virtues to AI systems, assessing their capacity to meet recognized criteria. Throughout this examination, I critically engage with the conclusions proposed by Constantinescu and Crisp (2022), with a specific focus on the gap between value alignment and virtue attribution. In discussing this gap, I provide answers to the following questions: i. Is value alignment a necessary or sufficient condition for ascribing virtues; ii. Does value alignment provide solid ground to justify the use of AI-based systems and tools whose deployment hinges on appeals to justice?

Emphasis on Virtues in the Catholic Church's Reflection on the Use of AI

Roman Globokar

The Catholic Church is involved in various ways in the wider societal debate on the ethical use of AI. In this paper, we will present the views of some Catholic moral theologians (Benanti, Kirchschlaeger, Spadaro) with a focus on the ethics of discernment and the identification of virtues that are particularly relevant for the formation of personal conscience and collective social consciousness within the digital age. Catholic theologians base their reflections on the principles of social doctrine, from which the following seven fundamental attitudes can be extracted: 1. The centrality of the human person and respect for their intrinsic dignity, 2. Ensuring the common good (inclusiveness and special attention to the marginalised), 3. Justice (ensuring equal opportunities), 4. Solidarity, 5. Subsidiarity, 6. Integrity (honesty, transparency and ethical integrity), 7. Responsible stewardship (towards the natural environment and future generations).

On the initiative of the Pontifical Academy for Life, the Rome Call for AI ethics was launched to the public on 28 February 2020 and contains six fundamental principles: transparency, inclusion, responsibility, impartiality, reliability, security and privacy.

In our contribution, we will pay particular attention to an analysis of Pope Francis' message on the occasion of the World Day of Peace, 1 January 2024, entitled "Artificial Intelligence and Peace", in which the Pope reflects on the complex relationship between AI and the betterment of humanity. In the encyclical, the Pope highlights the following virtues: 1. human dignity and fraternity, 2. justice and common good, 3. transparency, security, equity, privacy and reliability, 4. responsibility, 5. peaceful and fraternal coexistence, 6. ethical development of algorithms, 7. education for critical thinking and responsible use, 8. strengthening international law and cooperation.

Based on the literature studied, we will synthesise and identify some of the fundamental virtues that are rooted in a Christian view of human beings and that should guide the proper use of artificial intelligence in various areas of personal and social life.

LIST OF PARTICIPANTS

1. Eugen Dolezal[1], University of Graz, *M.Th.,Assistant Professor (pre doc), University of Graz, Department of Ethics and Social Teaching, Faculty of Catholic Theology*
2. Michael Paskaru, Tilburg University, York University, michaelpaskaru@gmail.com, *M.Phil., Teacher Assistant, Ontario College of Arts & Design*
3. Mateja Centa Strahovnik, University of Ljubljana, *Ph.D., Assistant Professor and Research Fellow, Faculty of Theology, University of Ljubljana, Centre for Human-Centered Artificial Intelligence and the Ethics of New Technologies, Institute of Bioethics*
4. Vojko Strahovnik, University of Ljubljana, *Ph.D., Professor, Faculty of Arts & Faculty of Theology, Centre for Human-Centered Artificial Intelligence and the Ethics of New Technologies*
5. Diana Daly, University of Arizona, *Ph.D., Associate Dean, Undergraduate Academic Affairs and Student Success Associate Professor of Practice, University of Arizona*
6. Noreen van Elk, University of Vienna, Noreen. *Ph.D., University Assistant (Postdoc), University of Vienna, Department of Systematic Theology and Ethics*
7. Ivan Cerovac, University of Rijeka, *Ph.D., Assistant Professor, Department of Philosophy, University of Rijeka*
8. Dušan Rebolj, University College London, *Ph.D. Candidate in Political Theory, University College London, Department of Political Science & School of Public Policy*
9. Tomislav Furlanis, University of Rijeka, *Ph.D., Laboratory for Ethical Aspects of Advanced Digital Technology, University of Rijeka*
10. Łukasz Białkowski, University of the National Education Commission in Cracow, *Ph.D., Assistant Professor at Department of Art Studies at the Institute of Art and Design, University of the National Education Commission, Cracow*
11. José Antonio Pérez-Escobar[2], University of Geneva, *Ph.D., Postdoctoral Researcher, University of Geneva*
12. Marko Robnik-Šikonja, University of Ljubljana, *Ph.D., Professor, University of Ljubljana, Faculty of Computer and Information Science*
13. Saša Poljak Lukek, University of Ljubljana, *Ph. D., Assistant Professor at the Department of Marriage and Family Therapy and Psychology and Sociology of Religion, Faculty of Theology, University of Ljubljana*
14. Jonas Miklavčič, University of Ljubljana, *Ph.D., Assistant, Faculty of Theology, University of Ljubljana, Institute of Bioethics*

[1] Co-author: Christoph Spöck, University of Graz, *MEd, MA, MA, Research Fellow, Institute for Ethics and Social Studies*

[2] Co-author: Deniz Sarikaya, University of Brussels, *Ph.D., Postdoctoral Researcher, Vrije Universiteit Brussel, Centre for Logic and Philosophy of Science*

15. Anita Lunić, University of Split, *Ph.D., Assistant, University of Split, Faculty of Humanities and Social Sciences*
16. Roman Globokar, University of Ljubljana, *Ph.D., Associate Professor, University of Ljubljana, Faculty of Theology*

TEOF | UNIVERSITY OF LJUBLJANA
Faculty of Theology

CENTRE FOR
HUMAN-CENTRED ARTIFICIAL INTELLIGENCE
AND THE ETHICS OF NEW TECHNOLOGIES