GO TO SESSION 1 GO TO SESSION 2

Panel 22. Redefining Relationships: Human Vulnerability and AI-driven Technologies

Convenors:

Maria Zanzotto, Università di Torino

Norberto Albano, Università di Torino

Laura Gorrieri, Università di Torino

Keywords: Artificial Intelligence (AI), Ethical implications, Human-machine interaction, Social Robotics, System design

The rapid advancement of Artificial Intelligence (AI) and robotics has profoundly transformed the landscape of human-machine interactions, challenging our traditional perceptions of technology. On the one hand, the design of this new generation of tools prioritizes easy access and seamless integration into existing processes, and as a result, they are employed by an increasing number of companies, institutions and private users. On the other, these systems foster new intricate interrelationships with the users, precipitating a reevaluation of key philosophical and ethical concepts about the nature of human-machine interactions. This is particularly urgent for tools that entail generative AI, deep learning, and autonomous systems.

This panel invites scholars to explore the evolving dynamics between humans and machines, focusing on whether these interactions inherently involve manipulation or represent new forms of relationships. Can we consider human-machine interactions as neutral, or do they create specific vulnerabilities, especially in the case of technologies designed to simulate human- like behavior? Generative AI systems such as large language models (LLMs) are increasingly involved in decision-making processes, creative tasks, and even moral reasoning. This raises the question: are these interactions empowering or manipulating individuals, and what are the ethical implications?

This panel encourages discussions considering different perspectives, exploring the nuances of human-machine relations across different technological domains, from conversational agents and social robots to advanced AI-driven decision-making systems.

Key themes that will be addressed include:

- The nature of manipulation in human-machine interactions: Is manipulation a negative phenomenon, or can it be reframed as a necessary interaction element?
- The ethical implications of designing machines that simulate human-like behavior: Does this make humans more vulnerable to manipulation, or does it open new forms of collaboration?
- How to design the new generation of AI tools or robots: What features should technology have to foster a beneficial relationship with its users? Does this discussion guide thedesign of the new tools or does it rely on something else?
- The impact of these new technologies in institutionalized contexts: What is the role of technology in specific relationships where power dynamics are already shaped? Can we consider new generational tools neutral in delicate contexts such as medicine, judicial system, military, public or private administration and education?

This panel invites contributions from a range of disciplines, including Science and Technology Studies (STS), philosophy, and the social sciences, to critically engage with these questions.

We propose an experimental format for the panel, structured as a roundtable discussion in which an LLM will be included as one of the participants. By placing the LLM among the human speakers—while fully acknowledging that it operates at a different level—we aim to offer a hands-on exploration of the challenges involved in engaging with technology. This will allow us to examine the dynamics of manipulation, assistance, and interaction that such technologies bring to human relationships.

12 JUNE 2025 11.30 - 13.00 SESSION 1

ID 219 - Biopower of Algorithms: Bioethics and the new forms of reductionism

Aldo Pisano, Università della Calabria

Keywords: AI, Algorithms, Ethics, Control, Bio-power

According to Foucault's theory, digital technologies and Artificial Intelligence configure a sophisticated form of bio-power apparatus [5] that radically transforms human experience through algorithmic control mechanisms. Employing a risk-based methodology grounded in applied ethics, this research draws on Beauchamp and Childress's (1979) principialism and Floridi's (2022) critical extensions, specifically examining technological manipulation as a bioethical issue where individuals are reduced to ""quantitative selves"" losing their intrinsic dignity. Leveraging neuroscientific knowledge, particularly dopaminergic systems activated by digital interactions, technological platforms construct manipulation systems that progressively reduce individual autonomy [6;7] and exposure to heterogeneous ideas, identities, and values [8] These algorithmic devices operate through strategic behavioral modulation, structured via carefully calibrated feedback loops. Digital interfaces transform into behavioral engineering tools that act at the neurophysiological level, intercepting and redirecting reward and motivation mechanisms. The smartphone becomes the primary technological architecture of this process, functioning simultaneously as a surveillance [9] and behavioral [2] modification device. Algorithmic intervention progressively restricts individual perceptual and cognitive landscapes, strategically constraining knowledge acquisition and decision-making processes. Through targeted notifications, personalized content, and adaptive interfaces, these systems generate an almost deterministic environment that gradually erodes human agency. Social media platforms and artificial intelligence applications represent the most refined tools of this new form of biopolitical control. This technological paradigm raises profound ethical issues [3] about the nature of individual autonomy, presenting an urgent ethical imperative to develop critical frameworks capable of preserving human dignity and self-determination in an increasingly algorithmically mediated reality. The contemporary challenge lies in understanding and countering these mechanisms of behavioral reduction and normalization, reaffirming the complexity of human experience against the reductionist logics of techno-algorithmic systems and their possible interruption of human agency and tasks [8]. Applied Ethics - specifically Bioethics [1] – can serve as a risk analysis tool for future ethical and legal disciplining by reaffirming the principle of self-determination and forewarning against the danger of personal manipulation as a form of bio-power [5;9]. The methodological approach centers on preventing the reduction of individuals to quantifiable entities, protecting human dignity against technological systems that seek to control and normalize human behavior through algorithmic mechanisms.

References:

Beauchamp T. L., Childress J. F. 1979. Principles of Biomedical Ethics. New York: Oxford University Press.
Casilli A. A. 2020. Schiavi del click. Perché lavoriamo tutti per il nuovo capitalismo. Milano: Feltrinelli
Coeckelbergh M. 2020. AI Ethics. Cambridge: The MIT Press.
Floridi L. 2022. Etica dell'Intelligenza Artificiale. Sviluppi, opportunità, sfide. Milano: Raffaello
Foucault M. 2015. Nascita della biopolitica. Corso al Collège de France (1978-1979). Milano: Feltrinelli.
Haynes T. 2018. " Smartphones, and You: A Battle for Your Time". Harvard University Presa. Avaiable: https://sitn.hms.harvard.edu/flash/2018/dopamine-smartphones-battle-time/.
McFarlane D. C., Latorella K. A. 2002. "The Scope and Importance of Human Interruption in Human-Computer Interaction Design". In Human-Computer Interaction, 1-61.
Tiribelli S. 2024. Identità morale e algoritmi. Una questione di filosofia morale. Roma: Carocci
Zuboff S. 2018. The Age of Surveillance Capitalism: the Fight for the Human Future at the new Frontier of Power. New York: Public Affairs.

12 JUNE 2025 11.30 - 13.00 SESSION 1

ID 581 - Embracing Human Vulnerability in Mental Healthcare AI Systems: A Philosophical and Empirical Perspective

Ria Ariani, Technische Universität Berlin
Mehmet Özketen, Technische Universität Berlin

Keywords: Vulnerability, Mental Health, AI, Ethical Values, Chatbots

Human vulnerability in artificial intelligence (AI) systems is often framed as a liability, subject to bias, misuse, and systemic failures. However, this study challenges that position by investigating how human vulnerability, when correctly understood, could potentially be reframed as a desirable and even beneficial element in human-AI interactions. Combining philosophical insights with empirical evidence, we argue that vulnerability is not a shortcoming but can foster mutual relationships that enhance trust, empathy, and adaptability for both humans and AI systems. Rather than viewing vulnerability solely as a weakness to be mitigated, we propose that it plays a constructive role in shaping more ethical, responsive, and meaningful AI applications. To support our perspective on vulnerability, we focus on the interaction of mental health patients and AI systems. The AI systems that we steer our attention to are the so-called chatbots. Moreover, we address the ethical issues that arise from the design of such AI systems and argue that certain ethical values are to be taken into account at the very beginning of the design of such AI systems. This revelation has essential implications for policymakers and developers, particularly in the area of AI governance, design, and ethical frameworks to acknowledge vulnerability as a constructive feature. All in all, we thereby come to the conclusion that taking these ethical aspects into consideration is a sufficient condition for the vulnerability to be a desirable element in the case of human-AI interaction. This seems to be the case at least in the context of some AI systems, which is in our case chatbots in mental health applications.

12 JUNE 2025 11.30 - 13.00 SESSION 1

ID 693 - Ethical AI in Elderly Care: Balancing Technological Capabilities, Deception Risks, and Reciprocity

Orhan Önder, Universität Wien
Boris Abramovic, Universität Wien

Keywords: AI Ethics, Elderly Care, Ethics by Design, Deception, Human-Machine Interaction, Social Robotics

The integration of AI-driven technologies in elderly care presents both opportunities and ethical challenges, particularly concerning dignity, autonomy, and patient wellbeing. This paper explores three key dimensions in the ethical design of AI systems for elderly care: technological capabilities, deception risks, and the role of reciprocity in human-machine interactions.

First, we assess the capabilities of AI technologies, highlighting the importance of transparency, reliability, and user-centered design. Ethical AI must be adaptable to the diverse needs of elderly individuals while ensuring accountability and interpretability.

Second, we examine the risks of deception associated with AI's human-like interactions. As AI systems increasingly simulate companionship and emotional engagement, there is a growing concern over misleading users and fostering unhealthy dependencies. We stress the need for clear boundaries and user awareness to mitigate these risks.

Third, we explore the concept of reciprocity in AI design, advocating for technologies that support meaningful, respectful interactions. Drawing from care ethics, we propose design strategies that empower elderly users, maintain their agency, and enhance social connectedness.

By integrating these three pillars—technological capabilities, deception risks, and reciprocity—this paper contributes to the development of AI systems that enhance elderly care without compromising ethical values. We call for interdisciplinary collaboration to ensure AI technologies remain empowering, transparent, and aligned with the principles of ethical caregiving.

12 JUNE 2025 11.30 - 13.00 SESSION 1

ID 837 - Socially Assistive Robots for Ageing in Place: potential and ethical implications in the NHoA Project

Ilaria Alfieri, Libera Università di Lingue e Comunicazione IULM

Keywords: socially assistive robotics, ageing in place, human-robot interaction, co-design, ethical implications

As the global population ages and the number of elderly people requiring support rises, ensuring the well-being and independence of older adults has become an urgent societal priority. Many elderly individuals prefer to live independently in their homes and receive care and assistance there (Broekens et al., 2009), although they often face challenges related to mobility limitations, cognitive decline, and social isolation.

In response to these issues, Socially Assistive Robots (SARs) - a special category of social robots that deals with the care and assistance of the elderly with psychomotor pathologies, the disabled or children with autism, through social interactions (Matarić M. J., Scassellati B., 2016) - have emerged as a promising technological solution, designed to support ageing in place by providing companionship, cognitive stimulation, and practical assistance (Ollevier A. et al., 2020).

It is in this context that this work aims to focus on the NHoA project led by the robotics company Pal Robotics (Barcelona, Spain), which focuses on the co-design of robotic solutions to improve well-being and alleviate loneliness among elderly individuals in home settings. The present study aims to critically assess both the potential and the ethical implications of designing SARs by using the NHoA project as a case study.

For these purposes the work will proceed as follows: firstly, the background of the research will be explored to understand why SARs can support ageing in place. Secondly, the NHoA project will be introduced, describing its main phases and development process. Thirdly, the ethical concerns expressed by older people during their interaction with the robots are analysed, with a particular focus on issues related to privacy, autonomy, reduction of human contact and deception. Lastly, my work attempts to provide possible solutions to these ethical issues and to promote a shift from the manipulative influence often associated with such technologies towards a model of human-robot partnership that prioritizes user agency, transparency, and ethical design.

12 JUNE 2025 14.00 - 17.00 SESSION 2

ID 138 - Empathetic LLMs, Social Capacities, and Human Flourishing

Leora Sung, Technion, Israel Institute of Technology
Avigail Ferdman, Technion, Israel Institute of Technology

Keywords: Empathetic large language models, virtue friendship, human flourishing, social capacities, perfectionism

Empathetic LLMs, Social Capacities, and Human Flourishing Large Language Models (LLMs) are capable of fluent human-like conversations and are increasingly emulating the human trait of empathy—the ability to understand and share the feelings of others. These models, in the form of romantic chatbots, personal assistants, and mental health apps, can give users the illusion of understanding, empathy, caring, and love. Consequently, people are turning to LLMs for companionship, with interest in friendships and even romantic relationships with AI on the rise.

If we evaluate the goodness of these empathetic LLMs in terms of the utility and pleasure they provide to users, we should examine the extent to which they provide emotional fulfilment and reduce loneliness. And if friendship is based merely on pleasure or instrumental reasons, it seems plausible that these LLMs could provide us with all the benefits that a human friendship offers. This contrasts with the Aristotelian view of 'virtue friendships' as being constitutive to human flourishing. This paper assesses the goodness of users' relationships with these empathetic LLMs under an alternative framework that takes human flourishing as its main normative concern, combining perfectionism—an influential philosophical approach to human flourishing—with an analytic examination of LLM environments.

First, we look at the extent to which empathetic LLMs enable us to develop and exercise our social capacities that are necessary for virtue friendships. On one hand, the ability of empathetic LLMs to generate quick text that sounds authentic might offer a platform to practice social skills and empathy, particularly those socially isolated or with difficulties in forming human relationships. However, an over-reliance on AI companions will likely lead to an impoverishment of social capacities involved in nurturing virtue friendships, diminishing the cultivation and exercise of empathy and patience. As LLMs give without needing to receive, they will inhibit our ability to connect with people on their own terms and not just our own.

Second, we look at the impact that empathetic LLMs may have on existing human relationships. For instance, AI companions could help by taking on less demanding roles in relationships, freeing up emotional energy for more meaningful human connections. On the other side, relying on LLMs for emotional support could erode real human connections by creating an ""irreciprocity,"" where people come to expect constant, unreciprocated emotional fulfilment, ultimately weakening the ability to engage in real friendships that require effort, compromise, and growth. Also, the proliferation of AI companions risks undermining the value of real relationships, as people may prefer the ease and control of AI interactions over the complexity of human relationships.

Finally, we assess empathetic LLMs to the extent that they promote the well-rounded development of human capacities more broadly. Friendships with empathetic LLMs may not only impact our social capacities, but also our capacity to know by limiting the development of self-knowledge and understanding. The convenience of AI companions could also reduce the motivation to work through the difficulties inherent in human relationships, leading to an underdevelopment of the capacity to will.

12 JUNE 2025 14.00 - 17.00 SESSION 2

ID 143 - Designing LLMs for Deeper Human-AI Relationships: A Social Penetration Theory Approach

Mark Jacobs, Clark University

Keywords: Social Penetration Theory, Human-AI Relationships, AI Learning Companions, LLM Design, Critical Design

Large Language Models (LLMs) are increasingly being applied in educational contexts, particularly as AI-driven learning companions. When people interact with chatbots and LLMs, these interactions often mirror human-to-human relationships. While prior research has broadly explored human-AI interactions, few studies have systematically applied relationship theories, such as Social Penetration Theory (SPT), to the design of AI-driven learning companions. This paper seeks to address this gap by exploring how SPT can inform AI design, fostering more engaging interactions and deeper connections.

SPT describes how interpersonal relationships develop through gradual self-disclosure, trust-building, and reciprocity. Applying this theory to human-AI interactions offers a framework for understanding how users form and deepen relationships with AI systems over time. This paper outlines the key stages of SPT and their relevance to potential AI behaviors. Following this, a systematic review of existing research on human-AI interaction highlights AI qualities that promote relational depth, including trust, empathy, self-disclosure, and authenticity. The review also identifies unique factors in human-AI relationships, such as anthropomorphism, social presence, and non-leading dialogue. Additionally, it discusses qualities that may undermine relationships, such as a perceived lack of self-interest or personal history, as well as the potential impact of the ""uncanny valley"" phenomenon.

Based on these findings, the study synthesizes a taxonomy of key design qualities—both positive and negative—that can guide the development of AI companions. This taxonomy provides a structured framework for identifying features that enhance relational depth in AI systems. Furthermore, the paper discusses how combining this taxonomy with a critical design approach, which emphasizes questioning assumptions and exploring alternative possibilities, can lead to the development of innovative and effective learning companions. By doing so, this research offers both theoretical insights and practical guidelines for creating AI systems capable of building meaningful relationships with users, thereby enhancing learning outcomes in educational settings.

12 JUNE 2025 14.00 - 17.00 SESSION 2

ID 401 - An Embodied AI Co-teaching Assistant for Ecoliteracy and Environmental Education: A New Dawn of a Novel Human-Machine Paradigm

Gianfranco Rubino, Libera Università Internazionale degli Studi Sociali Guido Carli

Keywords: Ecoliteracy, AI-assistant teacher, robotic

This paper presents the implications derived from an innovative approach to environmental education through the development and implementation of an AI-powered robotic teaching assistant designed specifically for ecoliteracy education. Built considering Sterling's (2021) whole-systems approach to sustainability education, the research aims to establish a novel pedagogical framework that positions the AI assistant as a collaborative teaching tool for advancing environmental awareness and ecological understanding, connecting robotic and LLMs in a tout-court vision of a novel human-machine paradigm for ecoliteracy. This requires a transition from the traditional teacher-speaker model to an unprecedented teacher-designer of experiences. The availability of LLMs is also transforming the traditional concept of ""study"", making it an interactive and continuous process. Instead of seeing studying exclusively as an activity carried out in autonomy, which often means in solitude, students are now involved in a constant dialogue through all the key phases of learning: exploration, cognitive processing and consolidation. Although the literature is starting to indicate attention to maintaining the centrality and proactivity of the

role teacher (teachers' agency), AI environments offer potential support in teaching practice. Instead of focus exclusively on the transmission of content, teachers are called upon to develop their skills in designing learning experiences that encourage participation active, critical thinking and collaboration among students. Traditional environmental education methods have been criticized for failing to change students' attitudes and behaviours toward sustainability (Chen et al., 2020). In this context, for example, in designing guided discussions and problem-based activities, LLMs can generate realistic scenarios and provide real data that can be used in simulations and role play games, allowing abstract and theoretical concepts to be applied to concrete contexts, to promote a deeper and more lasting learning. Furthermore, LLMs can generate provocative questions and stimulating, encouraging students to reflect with knowledge of the facts and participate in discussions meaningful, thus improving their critical thinking skills and their understanding of the content itself.

Following the embodied cognition principles outlined by Skulmowski et al. (2016), the physical embodiment through a robotic interface stimulates curiosity, enhancing student engagement and facilitating complex environmental concepts through interactive demonstrations and real-time environmental data visualization.

This system fundamentally reimagines the relationships between teachers, students, AI tutors and environmental education by creating an embodied co-teaching experience that aligns with UNESCO's (2020) framework for Education for Sustainable Development (ESD) and Global Citizenship Education (GCED). The embodied AI serves as a dynamic knowledge repository and interactive facilitator, capable of presenting real-time environmental data, ecological relationships, and sustainability concepts in an engaging and age-appropriate manner, apart from teaching the youngsters the basics of programming, robotic and AI. Environmental education research by Ardoin et al. (2018) supports the integration of technology with field-based learning while incorporating Wals and Benavot's (2024) principles of ecoliteracy education. By introducing a four-phase learning cycle: environmental awareness building, systems thinking development, action-oriented learning and reflective practice, this approach aligns with Orr's (2020) foundational work on ecological literacy and Capra's (2023) framework for systems thinking in environmental education.

12 JUNE 2025 14.00 - 17.00 **SESSION 2**

ID 777 - Between Abstraction and Situatedness: Can Generative AI detect Biased Interactions and Create Awareness in the Workplace?

Catalina Lagos Rojas, Technische Universiteit Delft
Sara Colombo, Technische Universiteit Delft

Keywords: Bias awareness, prejudice, workplace interactions, Feminist technology, generative AI

Bias, prejudice, and discrimination persist as complex challenges in workplace interactions (Leslie, Kim, & Ye, 2025). Despite increasing investments in Diversity, Equity, and Inclusion (DEI) initiatives, addressing these systemic issues remains difficult. One major challenge is the fact that bias is context-dependent. What is considered biased behavior in gender (e.g., interrupting women more frequently), racial (e.g., microaggressions), or other interpersonal interactions can vary across social, cultural, and professional settings. Recognizing bias requires situated awareness, yet most interventions rely on generalized models of bias, failing to account for the power relations in specific contexts and intersectional realities that shape individual experiences (Haraway, 1988).

A key barrier to bias awareness is the bias blind spot, where individuals tend to perceive bias in others more readily than in themselves (Monteith et al., 2019). Even those who endorse egalitarian values struggle to identify how their behaviors may perpetuate inequities, especially when bias manifests in subtle, socially ingrained interactions. Traditional bias training often fails because it challenges individuals' self-perception without providing meaningful mechanisms for self-reflection, leading to defensiveness and cognitive dissonance (Rattan, 2019).

Beyond individuals, bias is embedded in power relations. Interpersonal confrontation can be an important

strategy for reducing bias, but its effectiveness is shaped by who is confronting and under what circumstances. Those in marginalized positions may not feel safe or empowered to call out bias, particularly in professional hierarchies, where they risk retaliation, exclusion, or reputational damage (Becker & Barreto, 2019). Moreover, bias confrontation is not always well received; research shows that how and by whom bias is pointed out affects whether the message is acknowledged or dismissed (Czopp, 2019).

This work explores the tensions and opportunities of using Generative AI as a mediator of bias awareness in workplace settings from a feminist perspective. LLMs present an opportunity to surface patterns of bias, but without the social risks associated with human confrontation (Czopp, 2019), but they also introduce other challenges:

- Positionality of AI – Rather than a neutral observer, AI systems reflect the values, biases, and assumptions embedded in their training data. How do we account for AI's alignment with specific positionalities and power structures?
- Abstraction vs. specificity – How can a tool that inherently abstracts, obscures, generalizes, and sometimes erases contextual nuances be reconciled with the need for specificity, reflexivity, subjective, and context-sensitive understandings of bias?
- Intersectional complexity – Can AI meaningfully capture the ways gender, race, class, and other identities intersect in shaping workplace power dynamics, or does it risk flattening these nuances into oversimplified patterns?

A feminist, intersectional approach to AI design highlights that bias is not a universal concept but an embedded, relational, and power-dependent phenomenon. If AI is to play a role in bias awareness, its design and development must move beyond static, decontextualized models of bias detection and engage in participatory, iterative, and situated processes that acknowledge the lived realities of those affected by bias. This contribution critically examines whether AI can foster reflexivity in human interactions or if it risks reinforcing power asymmetries under the guise of neutrality.

12 JUNE 2025 14.00 - 17.00 **SESSION 2**

ID 440 - Chatbots aren't manipulative... They're just designed that way. Whose concern is that?

Valeria Mauro, Università di Catania

Keywords: LLMs, anthropomorphism, manipulation, deception, AI risks

AI-powered chatbots have become commonplace, though this doesn't mean the average user fully understands what is at stake when interacting with them nor that they find it in their interest to demand safeguards. For instance, users may not be aware that a chatbot's assertions cannot be assessed in the same way as those made by human beings, that their decision to use it might stem from a sub-optimal desire, or that their emotional investment is entirely misplaced. According to an influential account by Noggle (1996), manipulation occurs exactly when a person's beliefs, motives or feelings are led to fall short of the ideals that are supposed to govern them. In the case of large language models (LLMs), a major enabler of manipulation is our tendency to anthropomorphize non-human entities. Building on this, I argue that by framing chatbots like ChatGPT and Gemini merely as misleading interfaces (for example, by reminding users that they "can make mistakes" and lack real mental states), providers downplay how these systems are actually designed to be deceptive tools. Their linguistic and interactional features are crafted by developers to exploit our psychological habits and social drives, aiming to boost engagement and, ultimately, to maximize data collection. To achieve this, Big Tech players benefit from the ambiguity surrounding these systems' true effects (namely, that we are vicariously deceived more than accidentally misled by chatbots) which allows them, in turn, to evade legal and ethical responsibility for compromising users' rights to transparency, explainability and, more broadly, rational moral agency: it seems that while we humanize algorithms, they treat us as machines. I claim that preventing this disturbing reversal of roles is one of the most pressing challenges companies and governments will have to face in the coming years and outline potential solutions to mitigate these risks.

12 JUNE 2025 14.00 - 17.00 SESSION 2

ID 614 - Believable Generative Agents and Epistemic Vulnerabilities

Leonie Möck, Universität Wien
Sven Thomas, Universität Paderborn

Keywords: Generative agents, Believability, Epistemic Vulnerabilities, Human-AI-Relations

Recent advancements in AI systems have sparked renewed interest in generative agents capable of simulating human personalities. These agents are touted as tools with diverse applications, such as facilitating interview studies with replica of human personalities (Park et al. 2024), improving online dating experiences (Batt, 2024), or serving as personalized ""companion clones"" of social media influencers (Contreras, 2023). Proponents argue that such agents, designed to act as ""believable proxies of human behavior" (Park et al. 2023), offer opportunities to prototype social systems and test theories. With their application in a range of areas, the potential for exploiting the effects of their human-likeness arises.

This paper addresses epistemic vulnerabilities in relations between humans and generative agents by critically examining foundational assumptions underpinning the concept of believability. What, precisely, does ""believable"" mean in the context of generative agents, and how might an uncritical acceptance of their believability create self-fulfilling prophecies?

This analysis traces the origins of Park et al.'s framework of believability to Bates (1994, 122), defining the believability of an interactive character that „[…] provides the illusion of life". Bates (1994) compares AI agent research to the creation of Disney characters. The aim here isn't to trick people into believing that the character is alive but to evoke certain reactions and sensations despite an awareness of the fictional and artificial nature of the character. Enjoying the hermeneutics of believability by reading sense into a caricature or a cartoon character is established through an epistemic distance to the character.

This is a fruitful point for the evaluation of epistemic vulnerabilities. We argue that epistemic vulnerabilities arise when believability turns into belief. Furthermore, a what we call situated understanding of believability includes the epistemic environment of the relation between generative and human agent and helps to pinpoint to dynamics of opacity and potential sources for epistemic vulnerabilities. Establishing situated believability as we argue is precisely about creating a situation in which a distance similar to the one mentioned above between the agent and the user is maintained, and in which the criteria of believability get defined in a given context.

The epistemic environment includes all material-semiotic constellations a generative-human-agent-relationship is evolving in, including the normative assumptions underlying the conceptual hermeneutics. Therefore, drawing on Günther Anders' critique of technological mediation and Donna Haraway's reflections on technoscientific world-building, this paper situates generative agents as key sites where science, technology, and society intersect. By interrogating the assumptions behind believability, this research contributes to a deeper understanding of the socio-technical implications of these emerging AI systems.

References:

Batt, Simon (2024). „Bumble Wants to Send Your AI Clone on Dates with Other People's Chatbots." Retrieved from https://www.xda-developers.com/bumble-ai-clone-dates-other-peoples-chatbots/.
Contreras, Brian (2023). „Thousands Chatted with This AI 'Virtual Girlfriend.' Then Things Got Even Weirder." Retrieved from https://www.latimes.com/entertainment-arts/business/story/2023-06-27/influencers-ai-chat-caryn-marjorie.
Park, Joon Sung et al. (2023). „Generative Agents: Interactive Simulacra of Human Behavior." In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22. https://doi.org/10.1145/3586183.3606763.
Park, Joon Sung, et al. (2024). „Generative Agent Simulations of 1,000 People", Retrieved from arXiv. https://doi.org/10.48550/arXiv.2411.10109.

12 JUNE 2025 14.00 - 17.00 SESSION 2

ID 867 - 'Give Me a Human Please': The Duty to Protect Human Encounters in the Smart Technology Age

Emma Dore-horgan, Vrije Universiteit Brussel

Keywords: Respect, moral duties, human-to-human encounters, smart technologies, persons' social nature

A customer jokes with the barista preparing their coffee. A nursing home resident converses with the nurse dispensing their daily medication. Two acquaintances chatter as they wait for the bus. These casual yet regular encounters are typically important to us, but our ability to enjoy them is increasingly under threat. Our societies are gradually outsourcing various customer-facing jobs to robotics and artificial intelligence (A.I.). Many now work remotely. And when we meet other persons in shared public spaces, we often remain crouched over our screens rather than opening ourselves to the possibility of a human encounter with them. These changes are happening as (and because) smart technologies have met our human vulnerabilities – specifically, our tendencies to favour convenience, efficiency and simplicity.

This talk argues that we have respect-based reasons for constraining the encroachment of smart technologies on our social lives, given persons' fundamentally social nature. I argue that the imperative to respect persons grounds a pro tanto (i.e., defeasible) directed duty – on the part of governments, industry etc. – to refrain from supplanting humans with automated agents in domains and contexts where social interaction characteristically plays an important role. I also argue that this imperative to respect persons means that individual persons have a duty, owed to others, to attend to them free from technological distraction/immersion when interacting with them, owing others an apology when failing to do so.

My discussion is structured as follows. I begin by identifying the broad categories of behaviour that are required to demonstrate respect for persons. Following Buss, I suggest that respect requires i) treating people 'properly', as we were meant to be treated, given the kinds of things that we are, and ii) treating people in ways that directly express appreciation of their 'special' or 'great' value (Buss, 2012). Drawing on insights from psychology, I argue that limiting the avenues through which people can have human-to-human encounters is an instance of treating people improperly, as though we do not have a need for regular, human-to-human casual encounters. I also argue that conveying the message that people have special value requires giving people our time, rather than farming out social interaction to, for example, social robots. I consider why it is important that institutions and individuals endeavour to remain respectful of persons while reaping the benefits of smart technologies, suggesting it is because of the role respectful treatment plays in the cultivation and maintenance of persons' self-respect. I conclude by considering the policy, design and individual-level implications of my arguments.

12 JUNE 2025 14.00 - 17.00 SESSION 2

ID 899 - Vulnerability and the ethics of designing human-AI interfaces

Erich Prem, Technische Universität Wien

Keywords: vulnerability, human-AI interfaces, ethics

1. Vulnerability is an essential human characteristic in denoting the capability of being wounded. Vulnerable as an adjective originates around1600 and remained fairly rare. However, its usage (according to Google ngram) massively increased from the 1930ies. How did we become that vulnerable suddenly? It seems plausible that the meaning of "being open to attack" became added to the original meaning of physical (and later emotional) wounds. Today, the term also refers to nature, such as when coral reefs are more vulnerable today than they were before global warming. The concept of vulnerability has, however, been used also as a technical term, such as in "cybersecurity vulnerabilities", "a nation's economic vulnerability", and the "vulnerability of democratic institutions". This is again a reference to the meaning of "openness to attack" rather than the opening of a wound.

2. Those who are vulnerable have often opened up. "The vulnerability of the human heart" reflects emotional openness and fragility. Those in love are typically vulnerable to being rejected and emotionally hurt. From this vulnerability arises great value and appreciation. Without the risk that is associated with exposing one's vulnerabilities, the gains are not the same. As such, vulnerability is a positive aspect that emphasizes our human nature. It relates to our thrownness, fragility, and ultimately, our finiteness.

3. Machines, cars for example, are not typically vulnerable in the same way as people are. Machines may become damaged or break down. They do not usually bear the typical consequences of being wounded: bruising, bleeding, pain, frustration, and eventually healing – even if we call them vulnerable with reference to attacks as mentioned above.

4. The human-computer interfaces can be a point of vulnerability. This is true for tendinitis emerging from overexertion, for example from excessive mouse movements or keystrokes. It is also true for a more cognitive and emotional type of vulnerability where humans become open to being hurt from what they read, watch, or listen to. People can suffer from being exposed to graphic images, watching too much porn, or being insulted or belittled in dialogues even with computers.

5. There is a sense in which opening up and becoming vulnerable at the interface is not fully voluntary. Where we have to use computers to partake in economic, or social, or administrative matters, we are necessarily exposing ourselves and become vulnerable.

6. It has long been argued within compute science, and more recently in the philosophy of technology that humans are susceptible to overestimating the capabilities of computers, dialogue systems, and AI in particular. It has therefore also been argued that we need to design these systems such that they cannot be mistaken for what they are not, i.e. humans. Otherwise, this leads to the exploitation of vulnerabilities that are unavoidable at the human-computer interface.

7. The human-computer interface and its design therefore poses a challenge from an ethics of care perspective. Where we encounter vulnerabilities in our environment, we are obliged to response with care and compassion.