beijing workshop on the

PHILOSOPHY OF AI

京师人工智能哲学研讨会

2024. 11. 7

Main Building A802, Beijing Normal University

Beijing Workshop on the Philosophy of AI

京师人工智能哲学研讨会

Time

2024.11.7 (Thursday)

Venue

Main Building A802, Beijing Normal University

Host

School of Philosophy, Beijing Normal University
北京师范大学哲学学院
Center for Studies of Values and Culture, Beijing Normal University
北京师范大学价值与文化研究中心
International Research Center for Analytic Philosophy, Beijing Normal University
北京师范大学分析哲学国际研究中心

Co-hosts

Department of Philosophy, The University of Hong Kong
香港大学哲学系
Department of Philosophy and Religious Studies, Peking University
北京大学哲学系
Institute of Foreign Philosophy, Peking University
北京大学外国哲学研究所

Fund

This workshop is funded by the Research Funds for Excellent Interdisciplinary Research Groups at Beijing Normal University (1233300003), National Social Science Fund Project (23BZX120), and High-end Foreign Expert Project of the Ministry of Science of China (G2023111010L).

本次活动由北京师范大学优秀青年创新团队跨学科交叉项目（项目批准号：1233300003），国家社会科学基金项目（23BZX120）和国家级外专项目（G2023111010L）支持

Workshop Program

9:00 – 9:10	**Open Remark:** Hong Li Group Photo
	Chair: Chuang Ye
9:10 – 10:10	Xing Xie **Societal AI: Tackling AI Challenges with Social Science Insights**
10:20 – 11:20	Herman Cappelen **AI Survival Stories: A Taxonomic Analysis of AI Existential Risk (co-authored with Simon Goldstein, and John Hawthorne)**
11:20 – 13:00	*lunch break*
	Chair: Haiqiang Dai
13:00 – 14:00	Sebastian Sunday Grève **Can Machines Acquire Human Mindedness?**
14:05 – 15:05	Linus Huang **AI, Normality, and Oppressive Things**
15:10 – 16:10	Junqi Wang **Theory of Mind: from a Modelling Perspective**
16:15 – 18:00	**Roundtable Discussion** Chair: Yiwen Zhan Discussants: All Participants
18:00	*dinner*

Participants

Pierrick Bourrat (Macquarie University)

Herman Cappelen (The University of Hong Kong)

Long Chen 陈龙 (Beijing Normal University)

Haiqiang Dai 代海强 (Beijing Normal University)

Yifeng Ding 丁一峰 (Peking University)

Jianqiao Ge 葛鉴桥 (Peking University)

Zhiwei Gu 顾知巍 (Fudan University)

Linus Huang 黄大伦 (Hong Kong University of Science and Technology)

Xiaoyu Ke 柯晓宇 (Zhejiang University)

Hong Li 李红 (Beijing Normal University)

Jixin Liu 刘佶鑫 (Sichuan University)

Qiaoying Lu 陆俏颖 (Peking University)

Qianqian Sun 孙骞谦 (Central Academy of Fine Arts)

Sebastian Sunday Grève (Peking University)

Xinyuan Tian 田馨媛 (Berggruen Institute, China Center)

Junqi Wang 王俊淇 (Beijing Institute for General Artificial Intelligence)

Xing Xie 谢幸 (Microsoft Research Asia)

Chuang Ye 叶闯 (Shanxi University)

Yiwen Zhan 展翼文 (Beijing Normal University)

Ziheng Zhou 周自横 (University of California, Los Angeles)

Student Assistants

Zihan Huang 黄梓涵 (Beijing Normal University)

Xiao Wang 王骁 (Beijing Normal University)

Liang Xie 解亮 (Beijing Normal University)

Jiaocheng Zhang 张骄成 (Beijing Normal University)

Peng Zhang 张澎 (Beijing Normal University)

Bios

Pierrick Bourrat is a Senior Lecturer and DECRA Fellow at Macquarie University, Sydney, Australia. He specialise in philosophy of biology. Together with Paul Griffiths, he heads the Theory and Method in Biosciences group. His work focuses on core concepts of evolutionary theory, such as fitness and heritability, as well as evolutionary transitions in individuality. His other interests include causation, the interplay between biological and cultural evolution and the evolution of religious beliefs.

Herman Cappelen is a Chair Professor of Philosophy at the University of Hong Kong. He's a Director of the AI&Humanity-Lab at HKU and the Director of the MA Programme in AI, Ethics, and Society (at HKU). Prof. Cappelen is an elected member of the Academia Europaea and an elected member of the Norwegian Academy of Science and Letters. He is a leading expert in philosophical methodology, philosophy of AI, philosophy of language, and conceptual engineering. He is the author of eleven monographs and many influential papers.

Long Chen is currently a lecturer at the School of Philosophy, Beijing Normal University. He received a master's degree and a doctorate in philosophy from Peking University and King's College London respectively. His main research areas are philosophy of mathematics and philosophy of logic, especially Goedel's philosophy of mathematics and the problem of vagueness as well as the liar and related paradox and the problem of logical normativity. He also has a keen interest in the history of 20th-century analytical philosophy and formal philosophy in general.

Haiqiang Dai, Associate Professor, School of Philosophy at Beijing Normal University, General Secretary of Chinese Wittgenstein Society, research areas including Wittgenstein, philosophy of language, philosophy of normativity, metacognition, philosophy of dreams.

Yifeng Ding is an assistant professor at the Department of Philosophy at Peking University. He obtained his Ph.D. in Logic and the Methodology of Science from UC Berkeley's logic group under the supervision of Wesley Holliday in 2021. Before that, he received BA in philosophy and economics from Peking University in 2015. He works mainly in modal logic and axiomatic social choice theory. In modal logic, he has published works on logics for different kinds of knowledge, theories of non-normal modal logics, logics with propositional quantifiers, and comparative logics for probabilities and cardinalities. In social choice, his current interest is in axiomatic characterizations for margin-based voting methods, especially those that respond nicely to expansions of the pool of candidates and voters.

Jianqiao Ge is teaching at Academy for Advanced Interdisciplinary Studies (AAIS), Peking University. Her research interests are focused on scientific studies of brain intelligence and social cognition. She is the Principle Investigator/Co-Investigator of more than 10 research grants supported by Ministry of Science and Technology of China, National Natural Science Foundation of China, and Beijing Municipal Science & Technology Commission. She has published more than 20 referred research articles on leading academic journals such as Nature Neuroscience, PNAS, and awarded 2 national patents. She was awarded Berggruen Fellow 2021-2022.

Zhiwei Gu completed his PhD at CEU in 2020 and joined Fudan University as a postdoc in 2022, becoming an assistant professor in 2024. His research focuses on the philosophy of perception, mind, and metaphysics. He is currently working on a book that addresses challenges to naïve realism from cognitive science. He attempts to argue that low-level cognitive mechanisms can causally explain but not refute the high-level experiential relation.

Linus Huang is a Research Assistant Professor at the Division of Humanities and Centre for AI Research, Hong Kong University of Science and Technology, and a StarTrack Scholar at Microsoft Research Asia. His interdisciplinary research focuses on AI ethics, philosophy of cognitive science, and embodied cognition. His work tackles critical issues like reducing algorithmic bias, aligning AI with human values, and understanding human intelligence. Currently, he leads the funded project Engineering Equity, which explores AI's potential to mitigate implicit bias in HKUST, and another initiative on dynamic value alignment for large language models at Microsoft.

Xiaoyu Ke is an empirically-informed philosopher of mind and psychology working on topics related to the ethics and epistemology of emotion. Her interests also extend to emerging technologies related to emotions, such as affective brain-computer-interface technology. Xiaoyu received her PhD from the Philosophy-Neuroscience-Psychology program at the Washington University in St. Louis.

Hong Li is a Professor at the School of Philosophy, Beijing Normal University. She mainly works on analytic metaphysics, philosophy of language, and hermeneutics. Her representative works include the monograph: *The Convergence of Contemporary Western Analytical Philosophy and Hermeneutics*. She is the director of a Major Program of the National Social Science Fund of China on the philosophy of normativity.

Jixin Liu is an associate professor at the Department of Philosophy at Sichuan University. His research mainly focuses on modal logics, especially polyadic modal logics (where modal operators have more than one argument) and their applications (on knowledge, belief, evidence, and so on). He is also interested in how to characterize mathematical structures with (computable) modal languages. Other interests include social choice theory, the axiom of choice, metaphysics of time, and game theory.

Qiaoying Lu is an Assistant Professor in the Department of Philosophy at Peking University and a 2020-2021 Berggruen Institute Beijing fellow. Her research primarily focuses on the philosophy of biology and general philosophy of science. Key aspects of her work include establishing the theoretical basis for an extended gene-centered framework and utilizing structural causal models to examine genetic causality. She has also explored topics such as the units of natural selection, the resurgence of Lamarckian ideas, gene editing, and minimal cognition. Her recent research interests center on the evolution of biological cognition and consciousness and their relation to AI cognition.

Qianqian Sun is a lecturer at the School of Humanities, Central Academy of Fine Arts. He received his Ph.D. from the Institute of Foreign Philosophy at Peking University and completed postdoctoral research at the Institute of Logic and Cognition, Sun Yat-sen University. His research focuses on the philosophy of mind and philosophy of cognitive science within the analytic tradition. His current primary interests include action and perception, the mechanisms and architecture of cognition, and the fundamental units of cognition.

Sebastian Sunday Grève is an assistant professor in the Department of Philosophy and Religious Studies at Peking University. He is a Fellow of the Institute of Foreign Philosophy at Peking University and a former Berggruen China Fellow. He joined Peking University in 2019. Previously, he taught philosophy at the University of Oxford, where he gained his doctorate in 2018. In 2014, his essay 'The Importance of Understanding Each Other in Philosophy' was awarded the Annual Essay Prize of the British Royal Institute of Philosophy. Dr. Sunday works broadly in philosophy, on both practical and theoretical issues. He has published papers on topics ranging from aesthetics and the theory of knowledge to logic and the philosophy of mind, including artificial intelligence. Recent popular pieces include ‘AI’s First Philosopher’, ‘Can Machines Be Conscious?’, and ‘Nietzsche and the Machines’. Currently, Dr. Sunday is collaborating with colleagues in neuroscience and medicine at the University of Toronto on the use of large language models as a tool for clinical decision making as well as with colleagues from the Academy for Advanced Interdisciplinary Studies at Peking University on the perception of intelligence.

Xinyuan Tian is a senior program coordinator at the Berggruen Research Center, Peking University. She manages academic communications and organizes workshops and conferences on frontier science, technology, and philosophy. Xinyuan also leads interdisciplinary projects focused on creative futures.

Junqi Wang currently works as a researcher at State Key Laboratory of General Artificial Intelligence (Beijing Institute for General Artificial Intelligence, BIGAI). He got PhD in Mathematics at Rutgers University and Bachelor’s degree in Physics at Zhejiang University. He started working in AI area since working as Postdoc in the Cognitive and Data Science Lab at Rutgers University. Currently he is working on various topics, including the modeling and learning of value systems of human beings, social intelligence modeling and language model related topics. He has publications on various conferences and journals including NeurIPS, ICML, ICLR, CogSci, Entropy, etc. on various topics including Bayesian modeling theory, optimal transport, reinforcement learning, compression algorithm.

Xing Xie is a partner research manager at Microsoft Research Asia. He received his B.S. and Ph.D. in Computer Science from the University of Science and Technology of China in 1996 and 2001, respectively. Since joining Microsoft Research Asia in July 2001, Dr. Xie has focused on data mining, social computing, and responsible AI. His work has been recognized with several prestigious awards, including the IEEE MDM 2023 Test-of-Time Award, ACM SIGKDD 2022 Test-of-Time Award, ACM SIGKDD China 2021 Test-of-Time Award, ACM SIGSPATIAL 2020 10-Year Impact Award Honorable Mention, and ACM SIGSPATIAL 2019 10-Year Impact Award. He has delivered keynote speeches at notable conferences such as MDM 2019, ASONAM 2017, and W2GIS 2011. Dr. Xie serves on the editorial boards of ACM Transactions on Recommender Systems, ACM Transactions on Social Computing, ACM Transactions on Intelligent Systems and Technology, and CCF Transactions on Pervasive Computing and Interaction. He served as program co-chair of ACM Ubicomp 2011, PCC 2012, UIC 2015, SMP 2017, ACM SIGSPATIAL 2021, ACM SIGSPATIAL 2022, IEEE MDM 2022, PAKDD 2024, and IEEE BigData 2025. Dr. Xie is a Fellow of the ACM, IEEE, and China Computer Federation.

Chuang Ye is a Professor at the School of Philosophy, Shanxi University. He joined the faculty in 2022, and before that he taught at Peking University. He works at the intersection of philosophy of language and metaphysics. He has particular interests in fictional objects, virtual reality, empty names, and truthmaking.

Yiwen Zhan is a lecturer at the School of Philosophy, Beijing Normal University. He mainly works in metaphysics and epistemology. His research focuses on the philosophy of modality and formal epistemology. Recently, he is particularly interested in exploring the modelling of bounded rationality and processes of inquiry in epistemology and decision theory. He is also interested in metasemantics, and has a peculiar interest in the applications of plural logic in metaphysics.

Ziheng Zhou is currently a PhD Candidate at UCLA CS Department, co-advised by Demetri Terzopoulos and Song-Chun Zhu. His research interests lie in world model, ethics and value alignment in AI, or more intuitively, how can we enable AI to learn a causal abstract understanding of the world from perception, and how to understand and align to human ethics. Previously, he was a co-founding partner of an AI medical imaging startup VoxelCloud Inc., invested by Sequoia China, Tencent and other famous investing firms for about one hundred million dollars. He obtained a bachelor’s degree of Computer Science with two minors in Cognitive Science and Philosophy from UCLA. Aside from work and formal education, he also studied Chinese traditional philosophies, especially Buddhism, in depth along with a Buddhist monk scholar.

Abstracts

Title: AI Survival Stories: a Taxonomic Analysis of AI Existential Risk

Herman Cappelen, Simon Goldstein, and John Hawthorne

Abstract

Since the release of ChatGPT, there has been a lot of debate about whether AI systems pose an existential risk to humanity. This paper develops a general framework for thinking about the existential risk of AI systems. We analyze a two-premise argument that AI systems pose a threat to humanity. Premise one: AI systems will become extremely powerful. Premise two: if AI systems become extremely powerful, they will destroy humanity. We use these two premises to construct a taxonomy of 'survival stories', in which humanity survives into the far future. In each survival story, one of the two premises fails. Either scientific barriers prevent AI systems from becoming extremely powerful; or humanity bans research into AI systems, thereby preventing them from becoming extremely powerful; or extremely powerful AI systems do not destroy humanity, because their goals prevent them from doing so; or extremely powerful AI systems do not destroy humanity, because we can reliably detect and disable systems that have the goal of doing so. We argue that different survival stories face different challenges. We also argue that different survival stories motivate different responses to the threats from AI. Finally, we use our taxonomy to produce rough estimates of 'P(doom)', the probability that humanity will be destroyed by AI.

Title: AI, Normality, and Oppressive Things

Linus Huang

Abstract

While the pernicious biases of AI systems are widely acknowledged, much of the focus has been on overt expressions of these biases. In this paper, we shift attention to more covert ways in which AI systems contribute to oppression, using generative AI as a central case study. We identify and categorize how these systems perpetuate injustice through their content, style, and performance. Drawing on the literature on the social impact of technology, we argue that AI systems act as material anchors that reinforce oppressive normality, making them "oppressive things." Since oppression can manifest in varying degrees, we further analyze the factors that exacerbate the harmful impacts of AI systems. Recognizing the deep entanglement between AI and broader systems of oppression, we advocate for a more critical approach to AI design—one that considers its role in maintaining or challenging oppressive structures. Using Nancy Fraser's distinction between affirmative and transformative remedies to injustice, we conclude by critiquing existing measures and calling for a shift towards transformative solutions.

Title: Theory of Mind: from a Modelling Perspective

Junqi Wang

Abstract

Theory of mind (ToM) represents the capability of an agent (Human or AI agent) to understand other people by ascribing mental states to them. It is considered as fundamental building block of social intelligence and social behavior. Due to the abstractiveness of its definition, effect of theory of mind could be considered in almost all social activities, even in those that only "virtual" observers are considered in mind. In this talk, several approaches to modeling ToM will be introduced, mainly covering the Bayesian models. Examples in a simple Grid world will be introduced to demonstrate the corresponding models, introducing the iteration phenomenon which makes modelling ToM complicated. Finally, I will introduce some partial results dealing with iteration together with assumptions on how humans might use it.