Φ3

Workshop on Philosophy and AI

第三届哲学与人工智能研讨会

Time
2025.10.11–12 (Saturday–Sunday)

Venue
Art Station, 3rd Floor, Fudan Art Museum, Handan campus (there is a map below. You can't find "Fudan Art Museum" on popular maps, but you can search for "center for Japanese studies", and the art museum is nearby)

Host
School of Philosophy, Fudan University
复旦大学哲学学院
The Center for Philosophy and Science of Intelligence (PSI)
复旦大学智能科学与智能哲学研究中心

Co-Host
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（宗教学系）
School of Philosophy, Beijing Normal University
北京师范大学哲学学院
AI & Humanity Lab, The University of Hong Kong
香港大学人工智能与人文实验室
Institute of Foreign Philosophy, Peking University
北京大学外国哲学研究所
Hong Kong Ethics Lab, The University of Hong Kong
香港大学伦理实验室

Centre for Philosophy & the Future of Humanity, Peking University
北京大学哲学与人类未来研究中心
International Academic Exchange Committee, The Chinese Society for Philosophy of Nature, Science and Technology
中国自然辩证法研究会国际学术交流工作委员会

PHAI-3 Organizers

Boris Babic (The University of Hong Kong)

Herman Cappelen (The University of Hong Kong)

Zhiwei Gu (Fudan University)

Chuang Liu (Fudan University)

Yiwen Zhan (Beijing Normal University)

PHAI Program Committee

Boris Babic (The University of Hong Kong)

Herman Cappelen (The University of Hong Kong)

Haiqiang Dai (Beijing Normal University)

Jianqiao Ge (Peking University)

Zhiwei Gu (Fudan University)

Xiaoyu Ke (East China Normal University)

Chuang Liu (Fudan University; Chinese Academy of Sciences)

Jixin Liu (Sichuan University)

Qiaoying Lu (Peking University)

Nikolaj Jang Lee Linding Pedersen (Yonsei University)

Qianqian Sun (Central Academy of Fine Arts)

Sebastian Sunday Grève (Peking University)

Junqi Wang (Beijing Institute for General Artificial Intelligence)

Brian Wong (The University of Hong Kong)

Yiwen Zhan (Beijing Normal University)

Workshop Program

October 10, FRIDAY

14:00 – 24:00	Check-in
18:00	*dinner (TBA)*

DAY 1: October 11, SATURDAY

(Venue: Art Station, 3rd Floor, Fudan Art Museum, Handan campus)

9:30 – 9:50	**Workshop Opening** Group Photo
	Chair: Mingjun Zhang
9:50 – 11:00	Keynote: Zheng Zhang **LLM—The Split-Brainer That Talks Back**
11:00 – 11:10	*tea break*
11:10 – 12:00	Haoying Liu **Machine and the Pragmatics of the Mind**
12:00 – 13:30	*lunch break*

	Chair: Sebastian Sunday Grève
13:30 – 14:40	Keynote: Chaochao Lu **From Theory to Practice:** **A Framework for Understanding and Evaluating Consciousness in Large Language Models**
14:40 – 15:30	Qiaoying Lu **Why "Does AI Have Consciousness?" Is the Wrong Question: Lessons from the Evolution of Animal Consciousness**
15:30 – 15:45	*tea break*
	Chair: Nikolaj Jang Lee Linding Pedersen
15:45 – 16:35	Katsunori Miyahara **What Are We Asking when We Ask about Artificial Agency?**
16:35 – 17:25	Zhiwei Gu **A Transcendental Argument for AI Consciousness**
17:30 – 18:00	**PHAI business meeting**
18:30	*dinner*

DAY 2: October 12, SUNDAY

(Venue: Art Station, 3rd Floor, Fudan Art Museum, Handan campus)

	Chair: Jixin Liu
9:30 – 10:40	Keynote: Wangzhou Dai **From End-to-End to Step-by-Step**

10:40 – 10:50	*tea break*
10:50 – 11:40	Yuchuan Qiao **Towards the early detection of Alzheimer disease using AI-based MR imaging**
11:40 – 12:30	Keith Chan **Legal Liability for Unavoidable AI Harm: The Role of Explainability**
12:30 – 14:00	*lunch break*
	Chair: Jijiang Tian
14:00 – 14:50	Brian Wong **Large Language Model Injustice -- Structural, not Individual?**
14:50 – 15:40	Yue Zhu **Can We Reconcile AI Welfare and AI Safety?**
15:40 – 15:50	*tea break*
	Chair: Yuzhou Wang
15:50 – 16:40	Hua Shen **The Two-Way Street of Alignment: Measuring Values and Influence Between Humans and LLMs**
16:40 – 17:00	*Closing speeches*
18:00	*dinner*

Participants

Boris Babic (The University of Hong Kong)

Keith Chan (The Hong Kong University of Science and Technology)

Wangzhou Dai 戴望州 (Nanjing University)

Zhiwei Gu 顾知巍 (Fudan University)

Haoying Liu 刘皓滢 (Fudan University)

Jixing Liu 刘佶鑫 (Sichuan University)

Chaochao Lu 陆超超 (Shanghai AI Lab)

Qiaoying Lu 陆俏颖 (Peking University)

Katsunori Miyahara (Hokkaido University)

Yuchuan Qiao 乔豫川 (Fudan University)

Hua Shen 申华 (New York University Shanghai)

Brian Wong (The University of Hong Kong)

Zheng Zhang 张峥 (Amazon Web Services Shanghai AI Lab)

Yue Zhu 朱悦 (Tongji University)

Herman Cappelen (The University of Hong Kong)

Shiwei Chen 陈仕伟 (Tongji University)

Nikolaj Jang Lee Linding Pedersen (Yonsei University)

Chang Liu 刘畅 (Fudan University)

Chuang Liu 刘闯 (Fudan University)

Qianqian Sun 孙骞谦 (Central Academy of Fine Arts)

Xiao Tan 谭笑 (Capital Normal University)

Jijiang Tian 田继江 (Capital Normal University)

Shigeru Taguchi (Hokkaido University)

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

Kangyu Wang 王康予 (The University of Hong Kong)

Qiu Wang 王球 (Fudan University)

Yuzhou Wang 王昱洲 (Peking University)

Ruizhi Yang 杨睿之 (Fudan University)

Jie Yin 尹洁 (Fudan University)

Yiwen Zhan 展翼文 (Beijing Normal University)

Mingjun Zhang 张明君 (Fudan University)

Liqian Zhou 周理乾 (Shanghai Jiao Tong University)

Linfan Zhu 朱林蕃 (Fudan University)

Student Assistants

Zixun Jiang 江梓荀 (Fudan University)

Xiaodan Liu 刘晓丹 (Fudan University)

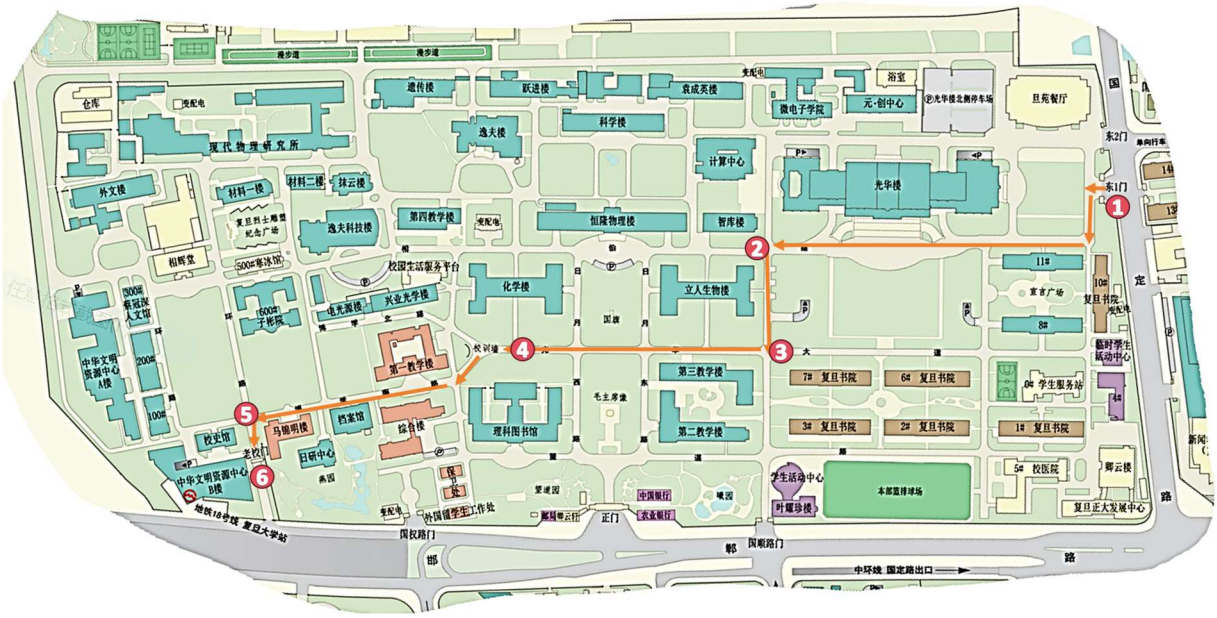
Shiqiang Lv 吕仕强 (Fudan University)

Qiuning Xiao 肖秋宁 (Fudan University)

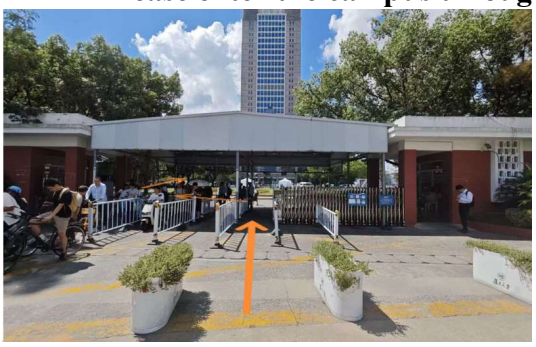
Kaiyao Xu 徐楷尧 (Fudan University)

Hanyao Zhang 张含瑶 (Fudan University)

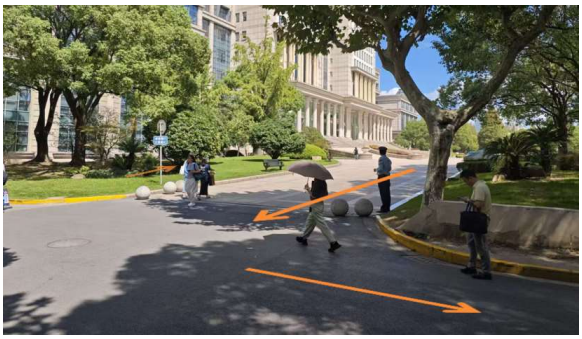
Map of Fudan University's Handan Campus



Please enter the campus through the East First Gate(No. 1).



No. 1 the East First Gate



No. 2 the Corner of Guanghua Building

No. 5 the Old Gate No. 6 Fudan Art Museum (Close to the Old Gate)

Abstracts

Title: LLM - The Split-Brainer That Talks Back

Zheng Zhang (Director at Amazon Web Services Shanghai AI Lab, Senior Principal Scientist)

Abstract: This talk presents a demystifying journey into Large Language Models (LLMs) through both technical exposition and philosophical inquiry. The first half provides an accessible crash course on LLM fundamentals: how next-token prediction transforms text into high-dimensional embeddings, how attention mechanisms enable context-aware pattern matching, and how the Transformer architecture orchestrates these components into seemingly intelligent behavior.

The second half reveals a fundamental paradox: LLMs exhibit "comprehension without competence." Drawing on recent empirical findings, I demonstrate how these models can perfectly explain algorithms they cannot execute, teach concepts they cannot apply, and articulate reasoning principles they cannot follow. This split-brain syndrome emerges from three architectural constraints: contaminated input representations that must serve multiple conflicting purposes, the mathematical impossibility of true computation within feedforward networks, and the absence of mechanisms to bridge instruction-following and execution pathways.

I argue that this paradox is not a bug but an inevitable feature of the current paradigm. LLMs are massive-scale, multi-level pattern completors that achieve remarkable fluency through sophisticated memorization and retrieval, not through genuine reasoning or computation. While they possess what we might call "general intelligence" (broad capability across domains), they fundamentally lack "generalizable intelligence" (the ability to discover rules and apply them systematically to novel problems).

The philosophical implications are profound: we have created entities that can discuss consciousness without experiencing it, explain logic without employing it, and teach mathematics without computing it. These split-brainers that talk back force us to reconsider fundamental questions about the nature of understanding, the relationship between language and thought, and what it means to "know" something. The talk concludes by examining what this tells us about both artificial and human intelligence, and why acknowledging these limitations is crucial for the responsible development and deployment of AI systems.

Title: Machine and the Pragmatics of the Mind

Haoying Liu (School of Philosophy, Fudan University)

Abstract: There is something puzzling about the issue of whether machine (or AI) can have a mind: even though it may approach human capacity, people tend to be grudging in attribution of mind or mental states to them. How to deal with this puzzle? It could seem that this is a question of the "mark of the mental", of finding a criterion or a set of criteria to tell whether something is mental or not. That is how many scholars in philosophy of mind are working on such issues. For example, Jonathan Birch and colleagues offer indicators of sentience to test whether animals that are evolutionary distant from human beings, such as cephalopods, have sentience. Patrick Butlin and colleagues have offered a list of indicators of consciousness based on the best current theories of consciousness to evaluate how likely some current Large Language Models could have consciousness.

Admittedly, it is helpful to figure out the features associated with mind or mental properties. However, in the case of machine we hardly have such clear markers. With animals and human beings, we seem to have clear markers. If their behaviors manifest certain patterns or if they perform certain functions, we are justified to judge them as having mind, psyche, etc. But these do not work for machines. No matter how complex the behavioral or functional patterns of a machine are, it seems that just in virtue of being a machine, its status of mentality must be subject to serious scrutiny. Thus, being machines almost disqualifies those machines entirely from having mind. Where markers are useful, it must be that the individuals in question are such that we may feel comfortable or reasonable to raise the possibility of their having mind. However, concerning the sense of puzzle about machine mentality, even the possibility of machine mentality is in question, much more than that of animal minds.

It is possible then that something ingrained in our current notion of mind and machine has created special hurdle for the idea of machine mentality, to the effect that many current discussions on machine mentality is systematically distorted. This possibility has been noted in many of Wittgenstein's observations on machine and mind, who poses the very possibility that questions regarding machine mentality are not "empirical" (Philosophical Investigation (PI) 360). Wittgenstein notes that mental terms such as "think", "sensation", etc. are said primarily "only of a living human being and what resembles (behaves like) a living human being" (PI 281-282). It is a matter of the "grammar" of mental terms that they are easily attributed to living beings, even simple ones like "a wriggling fly" (PI 284). Not only is the domain of mental terms restricted in our use, but the way they are used are also not entirely for cognitive purposes. As Wittgenstein observes, taking someone as having a mind is not an "opinion" but "an attitude towards a soul"; predicating "mind" of an individual might look like a plain statement, but actually it's not (only) making a statement, but doing something else. (PI, "Philosophy of

Psychology — A Fragment", iv 19-22) This extra linguistic work performed by the use of mental terms could be of a social nature. From Wittgenstein's account, we may observe that once one sees others as having mind, then one is in a state of social involvement. If one tries to suspend that perspective by imagining others as "automata, lack consciousness, even though they behave in the same way as usual", then "in the midst of [one's] ordinary intercourse with others," to the extent that one manages to conduct such imagination, one "will produce in [oneself] some kind of uncanny feeling, or something of the sort." (PI 420) Therefore, uses of mental terms may have the effect of building a social setting. In these observations and others, Wittgenstein offers insightful remarks on the pragmatics of discourse on mind and machine.

Contemporary philosophy of language does have resources to accommodate such pragmatic effects packed in seemingly unitary pieces of discourse. On one analysis of slurs, such expressions can perform more than one speech acts at once. Thus for example, using N-word to call someone is not only describing that person as belonging to a certain race, but also expresses endorsement of a derogating perspective toward the relevant social group (Camp 2018). Likewise, we may consider talking of some individual as "having mind" or "lacking mind" as having such rich pragmatic dimensions more than just statements. Attributions of mentality can also be regarded as ingrained with "perspectives", which are "modes of interpretation that structure an overall collection of thoughts in an intuitive, holistic way" (Camp 2018, p. 50). The perspective relevant to the use of "mind" and "machine" may serve some cognitive purposes, since regarding some individual as "mental" or "mechanic" offers clue on how one may proceed to understand or investigate it. But beyond that cognitive function, discourses of "mind" also has a practical aspect. Drawing from Wittgenstein's observations, we may say that this is a perspective of social involvement, with those "like us humans." Our recognition of them is not just a matter of "opinion" but "attitude," without which a sense of "uncanniness" would arise due to one's distancing from others. The word "machine" is used with the opposite of the perspective of "mind" or "mental", which is why the idea of "thinking machine" creates senses of puzzle.

Two lessons may be drawn from Wittgenstein's insights on the use of words like "mind" and "machine." First, such a pragmatic account may help explain the persistent rejection of machine mind or machine consciousness among some scholars. The intuition of some "gap" between mind and machines needs explanation, and an account of the pragmatics of "mind" and "machine" may offer such an explanation. More importantly, if such an account of the pragmatics is on the right track, then there is reason to think that the "mark of the mental" approach is wanting, for our concepts or terms of "mind" and "machine" have entirely different perspectives behind their uses, and we don't know yet how to integrate them. Real understanding of the possibility of machine mentality would then require a shift in such perspectives, which would not be a merely cognitive matter.

Title: From Theory to Practice: A Framework for Understanding and Evaluating Consciousness in Large Language Models

Chaochao Lu (Shanghai AI Lab)

Abstract: As large language models (LLMs) continue to advance, questions about their potential consciousness-like properties are moving from speculation to empirical study. This talk introduces a framework for defining, probing, and governing such phenomena. We propose a functional approach that translates key aspects of self-consciousness into measurable constructs. Using structured probing and causal interventions, we demonstrate how these capacities can be detected, manipulated, and in some cases enhanced through fine-tuning, suggesting that rudimentary forms of introspection already exist in today's models. Beyond empirical evidence, we survey theoretical accounts of machine consciousness and their computational relevance, while highlighting associated societal and safety risks. Different theoretical lenses shape how we interpret model behavior and how governance strategies should adapt to mitigate misuse and long-term risks. By integrating bottom-up experiments with top-down conceptual synthesis, this talk provides both methodological tools and policy insights for understanding and managing consciousness-like phenomena in LLMs.

Title: Why "Does AI Have Consciousness?" Is the Wrong Question: Lessons from the Evolution of Animal Consciousness

Qiaoying Lu (Department of Philosophy, Peking University)

Abstract: Much of today's debate about artificial intelligence turns on the question: Does AI have consciousness? Drawing on theories of the evolution of consciousness in animals, I argue that this question is poorly framed. Three major evolutionary approaches show why. First, if consciousness is simply cognition "from the inside," the real challenge is whether an artificial system can possess any "inside" perspective — something natural for embodied organisms but not obviously applicable to machines. Second, if consciousness depends on distinctive forms of sensory processing, as Prinz and others suggest, then AI systems trained on linguistic or digital data are not engaged in the perceptual pathways that shaped animal consciousness. Third, if consciousness is grounded in evaluative feelings that distinguish good from bad in the service of survival and reproduction, then systems without existential goals lack the foundations for such affective experience. Taken together, these perspectives indicate that asking whether AI has consciousness misapplies concepts rooted in biology. A more productive approach is to ask which specific

dimensions of animal consciousness are relevant to artificial systems, and what conceptual and practical consequences follow when some of those dimensions are realized.

Title: What are we asking when we ask about artificial agency?

Katsunori Miyahara (Hokkaido University)

Abstract: The question of whether an AI system can be an "agent" is a central concern in the philosophy of AI. However, this paper argues that the question "What is agency?" conflates at least four distinct questions: (1) The ontological question concerning the distinction between agents and non-agents; (2) The action-theoretic question of what distinguishes goal-directed action from mere movement; (3) The moral patiency question of what beings deserve moral consideration; and (4) The moral agency question of what constitutes a being that can be held responsible for its actions. These distinct questions are often "run together." While these facets of agency may be intrinsically linked, such connections cannot be assumed. By disentangling these four questions, I propose that we must avoid unreflectively running them together, and instead demand explicit justification for bridging claims between the different senses of agency to foster more precise and productive discussions about the status of artificial agents.

Title: A Transcendental Argument for AI Consciousness

Zhiwei Gu (School of Philosophy, Fudan University)

Abstract: This paper develops a transcendental argument for the existence of consciousness in artificial intelligence systems, drawing inspiration from P.F. Strawson's approach in "Freedom and Resentment." I argue that as AI systems become increasingly integrated into human society, the emergence of complex reciprocal attitudes between humans and AI constitutes a transcendental condition that necessitates the attribution of consciousness to these systems. Rather than treating AI consciousness as merely an epiphenomenal appearance or a convenient fiction, I contend that the reality of these intersubjective attitudes requires us to acknowledge genuine consciousness in sufficiently advanced AI. The argument suggests that consciousness is not merely a private, internal phenomenon but is constituted through patterns of mutual recognition and response that will inevitably develop between humans and artificial beings. This approach moves beyond both functionalist and substrate-based theories of consciousness to ground AI consciousness in the reality of our shared interpersonal space.

Title: From End-to-End to Step-by-Step

Wangzhou Dai （School of Intelligence Science and Technology, Nanjing University）

Abstract: Despite substantial advancements achieved by end-to-end learning architectures, these methods often struggle with tasks requiring explicit symbolic reasoning, robust generalization, and interpretability. Integrating statistical learning and symbolic reasoning remains a fundamental yet challenging goal in contemporary AI research. In this talk, we explore abductive learning as a principled framework to bridge neural models and formal logic, emphasizing the transition from purely end-to-end architectures toward structured, step-by-step reasoning. We will discuss how abductive reasoning—a logic-based mechanism to generate explanatory hypotheses—can guide machine learning models to autonomously discover symbols and causal relations from raw sensory inputs. Moreover, we will examine recent progress in abductive reinforcement learning, which recursively decomposes complex tasks into interpretable sub-tasks and discovers symbols and abstraction in open worlds. By moving beyond black-box approaches, this framework aims to improve model robustness, reduce data reliance, and enhance explainability, laying the foundation for the next generation of reliable and generalizable AI systems.

Title: Towards the early detection of Alzheimer disease using AI-based MR imaging

Yuchuan Qiao（Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University）

Abstract: The locus coeruleus (LC) is the first site of tau-related neurodegeneration in Alzheimer's disease (AD), but its microstructural signature remains elusive at the earliest, asymptomatic stage. We used cutting-edge distortion correction techniques to correct the distortions in brainstem region and investigate the LC integrity in cognitively intact individuals at increased AD risk. Voxel-wise and tract-specific analyses revealed localized reductions of fractional anisotropy and elevated radial diffusivity within the LC, preceding detectable cortical tau deposition or hippocampal atantrophy. These microstructural changes founded and the high-resolution LC-TEC atlas would support a non-invasive probe of preclinical LC pathology and open new avenues for early therapeutic targeting of the noradrenergic system in AD.

Title: Legal Liability for Unavoidable AI Harm: The Role of Explainability

Keith Chen (Institute for the Environment, The Hong Kong University of Science and Technology)

Abstract: According to the OECD, the number of AI-associated incidents attributed to the lack of explainability has skyrocket-ed after the release of ChatGPT in late 2022. Despite significant advancements in generative AI, it remains technically impossible to eradicate risks from rogue

behaviors such as specification gaming and, thus, challenging to determine who is liable. Current regulatory frameworks mainly focus on risk prevention but fail to address how liability should be allocated when harms do occur. In response to this critical gap in AI regulatory frameworks, we conducted a game-theoretic analysis to study the optimal design of legal liability for unavoidable AI harm under imperfect in-formation. Our analysis shows that the optimal liability regime dynamically hinges on the explainability of the AI system in question. Under this regime, courts can leverage industry-led explainability benchmarks to differentiate damages based on whether the developer can satisfactorily prove that the harm in question was unavoidable. Under this arrangement, developers are incentivized to invest in explainability techniques to reduce both the risk of their AI systems and their own liability exposure. Through this analysis, we demonstrate how a liability framework grounded in explainability can create economic incentives for developers to prioritize transparency and accountability.

Title: Can We Reconcile AI Welfare and AI Safety?

Yue Zhu (School of Law, Tongji University)

Abstract: In light of the significant recent progress in AI, particularly in the "Era of Experience," (Silver & Sutton, 2025) and in AI safety, specifically the Code of Practice for general-purpose AI models set forth in the European Union Artificial Intelligence Act, we re-examine the tension between AI safety and AI welfare as set out in Long, Sebo & Sims (2025). We conclude that it is becoming increasingly difficult to reconcile highly agentic AI welfare with rigid AI safety requirements.

Title: Large Language Model Injustice -- Structural, not Individual?

Brian Wong (Centre on Contemporary China and the World, The University of Hong Kong)

Abstract: From the generation of racist and sexist rhetoric, to manipulation and deception, the deleterious and unintended effects of large language models (LLMs) are increasingly ubiquitous. Who should bear responsibilities for such undesirable consequences, if any? This presentation will consider and reject two accounts of the normative stakes (or lack thereof) at hand - a) that LLMs' delivering such outputs should not be judged or appraised through ethical lenses of right/wrong, or b) that LLMs' outputs can be morally attributed to the actors who are involved in their training. The former account will be shown to be incompatible with our strongly held views and intuitions that reparative responsibilities can and ought to be assigned over these consequences - in a manner that cannot be strictly reduced to task responsibilities; the latter account will be shown to be far too demanding in evidentiary standards, as well as running afoul of key considerations pertaining

to fairness and traceability in the devising of applicable principles of justice. The preferred solution is one that invokes Iris Marion Youngian and post-Youngian works on structural injustice, which highlight the fact that individual and collective agents may accrue forward-looking responsibilities to redress the undesirable effects of LLMs, even if they are not blameworthy for such outcomes per se.

Title: The Two-Way Street of Alignment: Measuring Values and Influence Between Humans and LLMs

Hua Shen (Computer Science, NYU Shanghai)

Abstract: The integration of LLMs into society necessitates a broader understanding of alignment—one that accounts for bidirectional influence between humans and AI. This talk introduces a bidirectional human-AI alignment framework and presents a series of studies to understand and measure it. Our research operationalizes this concept through three critical lenses: 1) Value Misalignment: We introduce ValueCompass, a method to quantify contextual value alignment across cultures, and expose systematic value-action gaps in LLMs; 2) Perceptual Manipulation: We document user experiences with LLM dark patterns that manipulate belief and behavior; and 3) Dynamic Influence: We provide empirical evidence of bidirectional opinion dynamics in conversation, where both agent and human stances co-adapt. Together, our work provides new lenses to measure alignment, exposes critical risks, and charts a path for developing truly human-centered, responsible AI systems that are truly aligned through mutual understanding and adaptation.