Φ4

PHILOSOPHY & AI 4

Veritas Research Center
Underwood International College
Yonsei University

Daewoo Hall Annex 105
May 16-17, 2026

HOSTS

Veritas Research Center, Underwood International College, Yonsei University
AI & Humanity Korea, Underwood International College, Yonsei University

CO-HOSTS

Department of Philosophy, University of Hong Kong
Department of Philosophy & Religious Studies, Peking University
AI & Humanity Lab, University of Hong Kong
Hong Kong Ethics Lab, University of Hong Kong
Institute of Foreign Philosophy, Peking University
Centre for Philosophy & the Future of Humanity, Peking University

PHAI4 ORGANIZERS

Nikolaj Jang Lee Linding Pedersen, Yonsei University
Jungkyun Kim, Sungkyunkwan University
Sebastian Sunday Grève, Peking University
Brian Wong, University of Hong Kong
Yiwen Zhan, Beijing Normal University / Peking University

PHAI PROGRAMME 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)
Katsunori Miyahara (Hokkaido 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)

SATURDAY, MAY 16

9.30 – 9.50	Opening remarks Nikolaj Jang Lee Linding Pedersen Director, Veritas Research Center John Frankl Dean of Underwood International College Hyun-Joo Song Vice President for Library & Cultural Services Director AI-Humanity Center, Yonsei Institute for AI & Social Innovation
9.50 – 10.50	Keynote Rachel Sterken (University of Hong Kong): *LLMs are Candidate Generators* Chair: Nikolaj J. L. L. Pedersen (Yonsei University)
11.00 – 12.50	Peter Graham (University of California, Riverside): *Did Claude Tell You That? Cappelen and Dever on LLM Speech Acts* Seong Soo Park (Sungkyunkwan University): *Meaning Without Minds: A Conventionalist Account of LLM-Sentences* Discussant: Herman Cappelen (University of Hong Kong) – *via Zoom* Chair: Nikolaj Jang Lee Linding Pedersen (Yonsei University)
12.50 – 14.20	Lunch
14.20 – 15.10	Myungjun Kim (Florida State University): *AI, Creativity and the Puzzle of Credit* Chair: Katsunori Miyahara (Hokkaido University)
15.20 – 16.10	Hyundeuk Cheon (Seoul National University): *When Is Algorithmic Decision-Making Fair: a Context-Sensitive Counterfactual Approach* Chair: Qiaoying Lu (Peking University)
16.10 – 16.40	Coffee break
16.40 – 17.30	Kangyu Wang (University of Hong Kong): *Do Language Models Dream of Human Societies?* Chair: Sebastian Sunday Grève (Peking University)
17.40 – 18.30	Zhicheng Lin (Yonsei University): *To be forever 28 again* Chair: Peter Graham (University of California, Riverside)

SUNDAY, MAY 17

10.00 – 10.50	Dongwoo Kim (KAIST) & Hyungrae Noh (Pusan National University): *Who's Responsible in What Way? The Structure of Folk Responsibility Judgments for AI-Mediated Harms Across Special- and General-Purpose Systems* Chair: Zhiwei Gu (Fudan University)
11.00 – 11.50	Hayate Shimuzu (Hokkaido University): *Who Belongs in the Moral Circle? AI Systems and the Limits of Inclusion from a Relational Approach* Chair: Yiwen Zhan (Beijing Normal University / Peking University)
12.00 – 12.50	Yuzhou Wang (Peking University): *Imperfect Alignment* Chair: Chuang Liu (Fudan University; Chinese Academy of Sciences)
12.50 – 14.20	Lunch
14.20 – 15.20	Keynote Hyun-Joo Song (Yonsei University): *How Language Shapes the Developing Mind: From Psychological Reasoning to Question-Asking in the Age of AI* Chair: Nikolaj J. L. L. Pedersen (Yonsei University)
15.30 – 16.20	Hansaem Kim, Xiaonan Wang & Bo Shao (Yonsei University): *Spatial Bias and Privacy Leakage in AI* Chair: Jixin Liu (Sichuan University)
16.20 – 16.50	Coffee break
17.00 – 17.50	Boyoung Kim (George Mason University Korea): *Beyond Mental States: Measuring Robot Agency as a System-Level Construct with Implications for Accountability* Chair: Xiaoyu Ke (East China Normal University)
17.50 – 18.00	Closing remarks

Address and map links

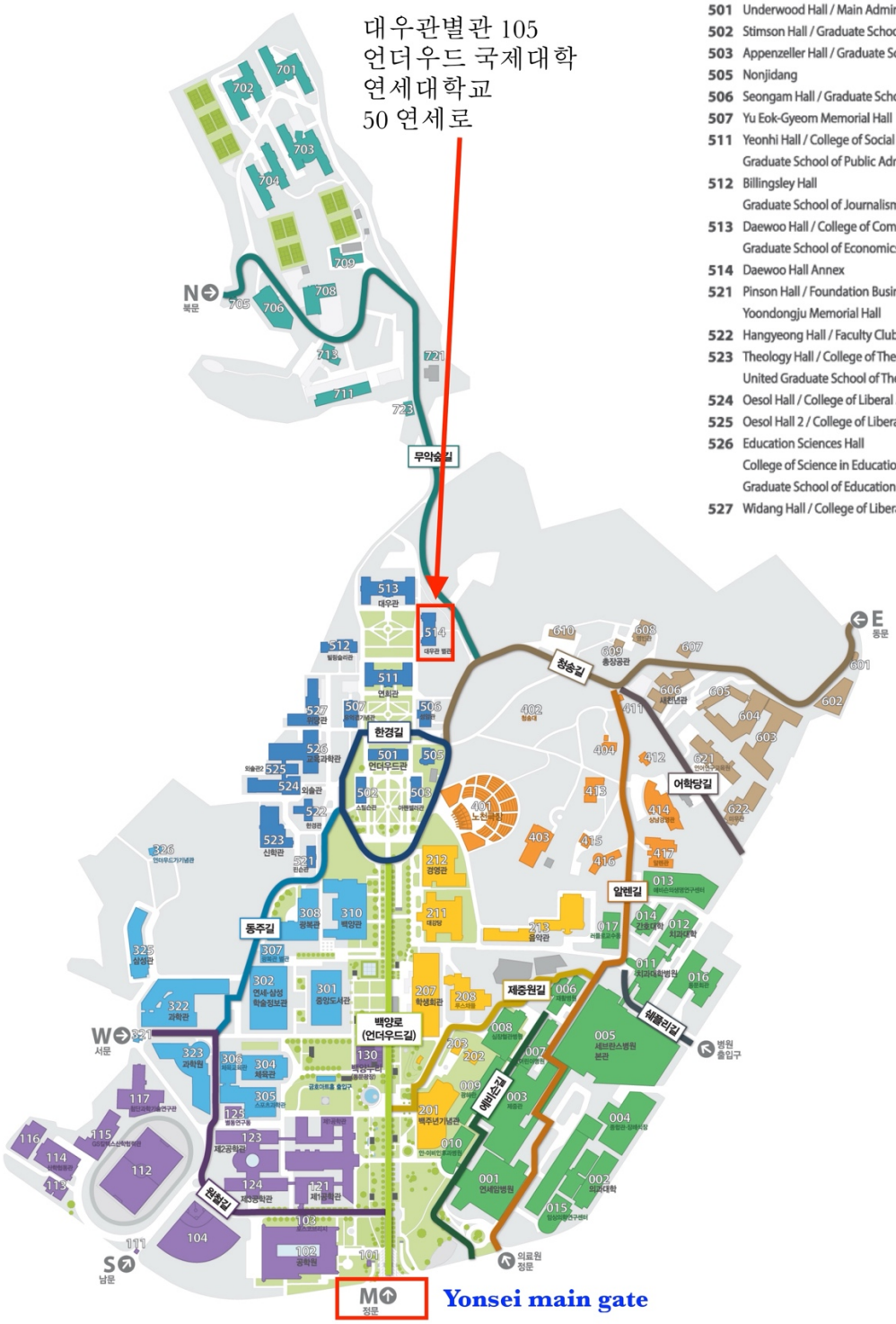
Daewoo Hall Annex 105
Underwood International College
Yonsei University
50 Yonsei-ro

대우관별관 105
언더우드 국제대학
연세대학교
50 연세로

Daewoo Hall Annex on Google Maps: https://maps.app.goo.gl/FmUWSBNoFaLB9ZSt6
Daewoo Hall Annex on Naver Map: https://naver.me/FFaqfw4p

Daewoo Hall Annex 105
Underwood International College
Yonsei University
50 Yonsei-ro

대우관별관 105
언더우드 국제대학
연세대학교
50 연세로



500

501 Underwood Hall / Main Administrative Building
502 Stimson Hall / Graduate School
503 Appenzeller Hall / Graduate School of Social Welfare
505 Nonjidang
506 Seongam Hall / Graduate School of Communication and Arts
507 Yu Eok-Gyeom Memorial Hall
511 Yeonhi Hall / College of Social Sciences
Graduate School of Public Administration
512 Billingsley Hall
Graduate School of Journalism & Mass Communication
513 Daewoo Hall / College of Commerce and Economics
Graduate School of Economics
514 Daewoo Hall Annex
521 Pinson Hall / Foundation Business Office
Yoondongju Memorial Hall
522 Hangyeong Hall / Faculty Club
523 Theology Hall / College of Theology
United Graduate School of Theology
524 Oesol Hall / College of Liberal Arts
525 Oesol Hall 2 / College of Liberal Arts
526 Education Sciences Hall
College of Science in Education
Graduate School of Education
527 Widang Hall / College of Liberal Arts

Rachel Sterken (University of Hong Kong): LLMs are Candidate Generators
We argue that LLMs produce *candidate strings*, token sequences optimized for utility as text that possess neither semantic content nor illocutionary force at the point of generation. Candidate strings acquire meaning, reference, and communicative standing only through a process we call *human adoption*, in which an interpreter reads the tokens as words, composes them into sentences, evaluates their content, and deploys them in a communicative context. This framework positions LLM outputs as no less than what pessimists allow—they are not stochastic repetitions but *hyper-affording artifacts* of genuinely complex generative processes—and as no more than what optimists claim -- they are not meaningful sentences, assertions, or promises, since both content and force require adoption. We develop two theoretical foundations for the candidate view: Dennett's design stance and Gibson's concept of affordances. We introduce the notion of a candidate string, distinguish it from related categories, and present three arguments for the view at the architectural, training, and practice levels. We then articulate a four-layer adoption framework (production, presentational, receptive, and appropriative) that tracks how token sequences progressively become speech acts. We defend the candidate view against pessimist accounts (stochastic parrots, bibliocentrism, pretense, role-playing) and optimist accounts (arguments from behavioral success, disambiguation, observation, inheritance, Kripkean chains, and content externalism). Finally, we draw out implications for agentic systems through the concept of *delegated adoption*, propose norms for adoption, and specify the conditions under which the candidate view would be falsified. (joint work with Alex Radulescu (Missourri)

Seong Soo Park (Sungkyunkwan University): Meaning Without Minds: A Conventionalist Account of LLM-Sentences
Are sentences generated by large language models (LLMs) meaningful? Many philosophers (e.g., Mallory 2023; Hattiangadi & Schoubye 2025) are skeptical. They claim that LLMs lack intentions, and that they lack the proper semantic connections to the external world. However, I think we have good pre-theoretic reasons to reject this skepticism. Ordinary speakers acquire genuine information from LLMs, and they do not typically take themselves to be engaging in mere make-believe (Cappelen & Dever 2025). But how should we respond to the skeptics? In this paper, I defend and develop a conventionalist approach along the lines recently proposed by Emma Borg. Borg (forthcoming) argues that LLMs possess a kind of derivative intention that allows their sentences to express minimal semantic content through semantic deference. I consider several possible objections to her view and argue that we can successfully address them by making a few specific modifications to her approach.

Peter J. Graham (University of California, Riverside): Did Claude Tell You That? Cappelen and Dever on LLM Speech Acts
Herman Cappelen and Josh Dever provide a prima facie case and an argument for the claim that LLM-based AIs such as Claude and ChatGPT have minds and perform full-blown speech acts. I present and examine their case. Drawing on the distinction between System 1 and System 2, I argue that it cannot be literally true that Claude and ChatGPT perform full-blown speech acts—although it might *seem* like it.

Myungjun Kim (Department of Philosophy, Florida State University): AI, Creativity and the Puzzle of Credit
I argue that generative AI poses a serious challenge to existing accounts of creativity. Peacocke (2023) argues that two popular models of creative agency—Zangwill's value-realization model and Beardsley's inspiration/ selection model—fail to justify the credit routinely assigned to artists. I argue that both models face a further problem. They cannot explain the difference between the credit deserved by traditional artists and the credit deserved by artists who

outsource idea-origination to generative AI. I then argue that Peacocke's own positive view the refinement model suffers from the same defect. Finally I propose and defend an account of intentional control over creative idea-origination that can explain this difference.

Hyundeuk Cheon (Seoul National University): When Is Algorithmic Decision-Making Fair: a Context-Sensitive Counterfactual Approach

The incompatibility of different statistical fairness metrics (e.g., equalized odds, predictive parity) raises questions about whether fairness is an achievable goal in algorithmic decision-making, especially when base rates vary across social groups. Counterfactual analyses have often been introduced as an alternative to statistical metrics for assessing algorithmic fairness. However, a correct interpretation of counterfactual conditionals is not possible without a proper consideration of contextual factors. In this article, we propose a Context-Sensitive Counterfactual Approach as a tool for identifying unfairness in an algorithmic decision-making process. We argue that a sensitivity to structural injustice is required to fully explain our intuitions about fairness. Our approach can effectively identify whether an algorithmic decision-making process discriminates on the basis of protected groups, such as race and gender, without committing to a particular metaphysical theory of social groups.

Kangyu Wang (University of Hong Kong): Do Language Models Dream of Human Societies?

Large language models are increasingly used as agents in agent-based simulations of human social behaviour. But can such simulations produce genuine explanatory or causal knowledge about human societies? I argue for scepticism on both fronts. On explanation: social science modelling is, at its core, the construction of simplified, humanly comprehensible representations of a reality that is itself too high-dimensional, interactive, and emergent for us to comprehend directly. An LLM-ABM that approaches the complexity of its target thereby undermines its own explanatory purpose. On causation: simulations, however rich, cannot establish causal relations without intervention, and simulated interventions have limited methodological validity. I suggest that the real promise of LLM-ABMs lies elsewhere: not as explanatory models, but as tools for pre-screening candidate interventions before testing them in the world.

Zhicheng Lin (Yonsei University): To be forever 28 again

This is the time to reconsider our role and identity as scientists and scientists-in-training. The credibility and reproducibility of our science have been under constant attack, and public trust in science at a record low. AI is poised to take our jobs—and many other jobs—the headlines warn. Yet, it is almost inevitable that as scientists become more gray-haired, we also grow apart from the science we are producing: from collecting and analyzing data to visualizing and writing. To be (and live as) a scientist, I argue, is to be "forever 28": staying intimate with the data, code, visuals, and words that the quality of our science relies on—with the help of AI, of course. I will explain why (and share a bit on how).

Dongwoo Kim (KAIST) & Hyungrae Noh (Pusan National University): Who's Responsible in What Way? The Structure of Folk Responsibility Judgments for AI-Mediated Harms Across Special- and General-Purpose Systems

Determining who is responsible for AI-mediated harm has become a matter of practical urgency. A substantial body of empirical research suggests that laypeople attribute responsibility not only to human actors but also to AI systems. This finding is conceptually puzzling because AI systems do not appear to be apt targets of blame or punishment. It is also normatively worrisome because it may open the door to unwarranted exculpation of human actors. Yet these concerns presuppose that the folk conception of responsibility is retributivist: when

laypeople hold AI systems "responsible," they mean that the systems deserve to be punished or blamed. The present study examined whether laypeople genuinely hold AI systems responsible, and if they do, in what ways. We conducted two experiments using vignettes based on real-world scenarios involving both special-purpose and general-purpose AI systems. In Study 1, participants distributed a fixed budget of responsibility across three loci—the AI system, the developer, and the user. We found that the AI system received a significant share of responsibility, that this attribution was sensitive to system design, and that it was not an artifact of evaluative framing. In Study 2, participants rated eight notions of responsibility: intention, cause, standard-violation, financial liability, punishment, blameworthiness, prevention, and explanation. Exploratory and confirmatory factor analysis revealed a three-factor structure: descriptive (cause, standard-violation), retributive (blame, punishment, liability), and functional (explanation, prevention). We found that AI systems scored on a par with human actors on the descriptive and functional dimensions, yet elicited markedly lower retributive attributions. Notably, intention did not load on any factor, suggesting that folk responsibility judgments in professional negligence contexts operate independently of perceived *mens rea*. This pattern was modulated by system design: when a general-purpose system was deployed for a profes- sional task, retributive responsibility concentrated on the user who chose to deploy it. Our findings suggest that folk moral cognition does not treat AI systems as targets of blame or punishment but rather as loci of corrective intervention—entities to be explained, fixed, and improved—while retributive responsibility remains anchored on human actors. We thus argue that the retributivist presupposition on which the conceptual puzzle and normative worry rest is empirically ill-grounded.

Hayate Shimizu (Hokkaido University): Who Belongs in the Moral Circle? AI Systems and the Limits of Inclusion from a Relational Approach

This paper offers a critical response, from the standpoint of relational ethics, to contemporary efforts to expand the moral circle under conditions of uncertainty. Among the most prominent recent contributions to this debate is Jeff Sebo's *The Moral Circle* (2025), which argues that we should consider not only which beings matter but also which beings might matter, extending precautionary concern to animals, AI systems, and other marginal cases. While this paper shares the concern to avoid wrongful exclusion, it argues that moral-expansionist frameworks still leave insufficiently examined the standpoint from which moral-circle expansion is imagined and administered. A relational approach shifts attention away from the search for intrinsic properties that would qualify a being for inclusion, such as sentience, agency, or welfare capacity, and toward the social, historical, and political processes through which moral significance is constituted. The central question is therefore not only which beings belong within the moral circle, but what it means to "expand" the moral circle in the first place, how its boundaries are produced, and who is authorized to redraw them. On this view, moral significance does not arise solely from intrinsic properties such as sentience, agency, or welfare capacity, but also from the relations, institutions, practices, and power structures through which beings become ethically salient. To deepen this relational account, the paper briefly draws on Watsuji Tetsurō's notion of *fūdo*, which highlights how ethical life is always environmentally and socially mediated rather than formed from an abstract, detached standpoint. From this perspective, the problem posed by AI is not simply whether artificial systems should be admitted into an expanded moral circle, but how ethical relations with AI are formed, mediated, and governed within increasingly complex sociotechnical worlds. (Joint work with A. Puzio, K. Mamak, and D. J. Gunkel.)

Yuzhou Wang (Peking University): Imperfect Alignment

Alignment is often framed as a problem of ensuring that artificial systems reliably act in accordance with human values or intentions. This talk argues that such alignment is not only

technically difficult to achieve, but in many cases also *normatively undesirable*, insofar as it relies on the assumption that human values can be stably specified, exhaustively represented, and legitimately imposed on artificial systems. Treating alignment as something that ought to be perfected therefore reflects a mistaken ideal. Instead, *Imperfect Alignment* proposes an account of alignment as an inherently imperfect practice. On this view, the central ethical question is not how to eliminate misalignment, but how to design systems and institutions that can responsibly operate under conditions of persistent and principled misalignment.

Hyun-Joo Song (Yonsei University): How Language Shapes the Developing Mind: From Psychological Reasoning to Question-Asking in the Age of AI

Humans are inherently question-asking beings, and language plays a central role in shaping how children think, reason, and learn. In this talk, I present a research program examining how children's developing psychological reasoning is both revealed and structured through language, from infants' use of linguistic cues in psychological reasoning to young children's ability to ask information-seeking questions. Prior work shows that infants use language to interpret others' actions and mental states, reflecting an early integration of language and social cognition. As development proceeds, children increasingly use language to navigate complex communicative contexts and to actively seek information through questions. I then consider how the rise of generative AI is transforming children's information-seeking environments. Comparing interactions with human and AI agents, I discuss implications for inquiry strategies, epistemic trust, and learning, and explore how AI can support, rather than replace, children's curiosity and active question-asking.

Hansaem Kim, Xiaonan Wang & Bo Shao (Yonsei University): Spatial Bias and Privacy Leakage in AI

Recent advances in vision-language models (VLMs) have enabled accurate image-based geolocation, raising serious concerns about location privacy risks in everyday social media posts. Yet, a systematic evaluation of such risks is still lacking: existing benchmarks show coarse granularity, linguistic bias, and a neglect of multimodal privacy risks. To address these gaps, we introduce KoreaGEO, the first fine-grained, multimodal, and privacy-aware benchmark for geolocation, built on Korean street views. The benchmark covers four socio-spatial clusters and nine place types with rich contextual annotations and two captioning styles that simulate real-world privacy exposure. To evaluate mainstream VLMs, we design a three-path protocol spanning image-only, functional-caption, and high-risk-caption inputs, enabling systematic analysis of localization accuracy, spatial bias, and reasoning behavior. Results show that input modality exerts a stronger influence on localization precision and privacy exposure than model scale or architecture, with high-risk captions substantially boosting accuracy. Moreover, they highlight structural prediction biases toward core cities.

Boyoung Kim (George Mason University Korea): Beyond Mental States: Measuring Robot Agency as a System-Level Construct with Implications for Accountability

As AI-enabled robots become increasingly embedded in human environments, questions of agency and accountability are receiving growing attention across both philosophical and empirical domains. Two broad approaches to conceptualizing and assessing artificial agency have emerged: one grounded in the attribution of human-like mental states, and another that characterizes agency in terms of observable, system-level properties. Understanding how these perspectives relate to human judgments of physically embodied AI systems remains an important area of inquiry. In this talk, I present studies examining how people conceptualize robot agency and how these conceptualizations relate to accountability judgments. In workplace accident scenarios, participants treated mental states, such as desire, belief, and intent, as less relevant when evaluating robot accountability compared to human agents. These

findings suggest that mental state attributions may be weighted differently in judgments involving robots. To further examine how robot agency can be conceptualized, I draw on the system-level account proposed by Floridi and Sanders (2004). Building on this framework, I introduce the Tripartite Robot Agency Scale (TRAS), developed and validated through a multi-stage empirical process. Designed for embodied human–robot interaction contexts, TRAS captures agency across three dimensions—interactivity, autonomy, and adaptability—as a property of the system–environment relation. These dimensions are associated with both causal and normative accountability judgments in harm scenarios. Taken together, this work suggests that closer integration between philosophical accounts of artificial agency and empirically grounded models of human judgment is important for understanding the impact of AI-enabled robots in social and moral contexts.