Topics in the Philosophy of AI

Hauptseminar, SoSe 2025

Konstantin Genin and Hong Yu Wong

konstantin.genin@gmail.com / hong-yu.wong@uni-tuebingen.de

12-14; Thursdays, Room X, Alte Burse

1. **April 17 - Searle's Chinese Room**

   Searle, John R. (1980) "Minds, brains, and programs." *Behavioral and brain sciences* 3.3: 417-424.

2. **April 24 - Skillful coping [Dreyfus / Dreydigger]**

   Dreyfus, Hubert. (1979). "From Micro-Worlds to Knowledge Representation: Al at an Impasse", *Mind Design II: Philosophy, Psychology, and Artificial Intelligence,* ed. John Haugeland, MIT Press, pp. 143-182.

   +

   Introduction to the MIT Press Edition of Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT press.

3. **May 8 - The Imitation game and the intentional stance [Dennett and Turing]**

   Turing, A.M. (1950) "Computing Machinery and Intelligence". Mind 49: 433-460.

   Dennett, D. (1981) "True Believers" (in Mind Design II)

   Jones, C.R. & Benjamin K. Bergen, B.K. (2025) Large Language Models Pass the Turing Test
   https://arxiv.org/abs/2503.23674 arXiv manuscript

   Wong, H. Y. (2025). Interrogating artificial agency. *Frontiers in Psychology*, *15*, 1449320.
   https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2024.1449320/full

4. **May 15 - Introduction to Philosophical Issues around Deep Learning [Milliere & Buckner]**

   Milliere, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models - Part I: Continuity With Classic Debates. *ArXiv, abs/2401.03910*.

https://arxiv.org/abs/2401.03910

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models-part ii: The way forward. *arXiv preprint arXiv:2405.03207*. https://arxiv.org/abs/2405.03207

Optional:
Buckner, Cameron. (2024) "What is Deep Learning, and How Should We Evaluate its Potential?" *From Deep Learning To Rational Machines.* Oxford University Press, pp. 48-93.

Levinstein, Ben. (2023) "A Conceptual Guide to Transformers Parts I-V" https://benlevinstein.substack.com/p/a-conceptual-guide-to-transformers

5. **May 22 - Continuation of Deep Learning + Probing for Representations [Hermann, Harding]**

Continuing with…
Milliere, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models - Part I: Continuity With Classic Debates. *ArXiv, abs/2401.03910*. https://arxiv.org/abs/2401.03910

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models-part ii: The way forward. *arXiv preprint arXiv:2405.03207*. https://arxiv.org/abs/2405.03207

And then:
Harding, Jacqueline. (2023) "Operationalising Representation in Natural Language Processing" *British Journal for the Philosophy of Science.* https://www.journals.uchicago.edu/doi/10.1086/728685

Levinstein, Benjamin A., and Daniel A. Herrmann. (2024) "Still no lie detector for language models: Probing empirical and conceptual roadblocks." *Philosophical Studies.* https://link.springer.com/article/10.1007/s11098-023-02094-3

Optional:

Herrmann, Daniel A., and Benjamin A. Levinstein. (2025) "Standards for belief representations in LLMs." *Minds and Machines* 35.1. https://link.springer.com/article/10.1007/s11023-024-09709-6

6. **June 4 Philosophy of AI workshop [9-16]**

Speakers:
Bojana Grujicic (Science of Intelligence, TU Berlin)
Daniel Herrmann (Philosophy, Groningen)
Michael Franke (Computational Linguistics, Tübingen)
Thomas Grote (ML, Tübingen)
Krisztina Orban (CIN/PONS, Tübingen)
Charles Rathkopf (Julich)

10-10.45 Charles Rathkopf
10.45-11.30 Thomas Grote
11.30-12.15 Bojana Grujicic

12.15-13.15 Lunch

13.15-14 Michael Franke
14-14.45 Krisztina Orban
14.45-15.30 Daniel Herrmann

15.30-16 Coffee Break

16-16.45 Poster session 1
16.45-17.30 Poster session 2

Language Models don't understand language (like you do): Emergent World Models & Causal Abstraction
Michael Franke (Computational Linguistics, Tübingen)

A popular claim in recent discussions of the abilities of generative AI in general, and language models in particular, is that such models must evolve veridical world models (of the causal data-generating process) in order to deliver the breath-taking performance we all got used to admire daily. I disagree and present an argument based on the task-optimality of emergent "world models" in generative AI. Concretely, focussing on autoregressive LMs, I argue that a model's inner representations are, if optimized for the task, most likely NOT veridical representations of the true data-generating process, but at best that of a *causal abstraction* of the true process (leaving out hierarchical structure of latent variables not directly reflected in the training data). Moreover, the ensuing representations are

most likely NOT of a causal nature, but merely representations of stochastic dependence.

7. **June 26 - Reinforcement Learning [Sutton, Haas, Butlin, Shea]**
   Butlin, P. (2024). Reinforcement learning and artificial agency. *Mind & Language*, *39*(1), 22-38.

   Haas, J. (2022). Reinforcement learning: A brief guide for philosophers of mind. *Philosophy Compass*, *17*(9), e12865.

   Haas, J. (2020). Moral gridworlds: a theoretical proposal for modeling artificial moral cognition. *Minds and Machines*, *30*(2), 219-246.

   Halina, Marta (2021). Insightful artificial intelligence. *Mind & Language*, *36*(2), 315-329.

8. **July 3 - AI consciousness**
   Chalmers, D. J. (2023). Could a large language model be conscious?. *arXiv preprint arXiv:2303.07103* https://arxiv.org/abs/2303.07103

   Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint arXiv:2308.08708* https://arxiv.org/abs/2308.08708

Requirements for Credit

***Shared requirements for all degrees and credit combinations:**

(a) Attendance (2 absences allowed without grounds)
(b) 1 Poster Presentation at workshop
(c) Complete all homework assignments (even if absent from class)

MSc ML / MSc Cogsci / MSc Computational Neuroscience

3 ECTS graded

* + 2000 Word essay on a topic to be determined in consultation with the instructors

BSc Cogsci

3 ECTS graded

* + 1500 Word essay on a topic to be determined in consultation with the instructors

MA Philosophy

3 ECTS ungraded

* + 1500 Word essay on a topic to be determined in consultation with the instructors

12 ECTS graded

* + 5000 Word essay on a topic to be determined in consultation with the instructors

MEd Philosophy/Ethics (2021 MHB)

3 ECTS ungraded

* + 1500 Word essay on a topic to be determined in consultation with the instructors

8 ECTS graded

* + 4000 Word essay on a topic to be determined in consultation with the instructors

MEd Philosophy/Ethics (2018 MHB)

3 ECTS ungraded + 5 ECTS graded

* + 4000 Word essay on a topic to be determined in consultation with the instructors

BA Philosophy (both 2021/2013 MHB) & BEd Philosophy/Ethics (2021 MHB)

6 ECTS graded

* + 3000 Word essay on a topic to be determined in consultation with the instructors

BEd Philosophy/Ethics (2015 MHB)

3 ECTS ungraded

* + 1500 Word essay on a topic to be determined in consultation with the instructors

3 ECTS ungraded + 6 ECTS graded

* + 4000 Word essay on a topic to be determined in consultation with the instructors

Other Tübingen Students

Please contact the instructors to make an arrangement. Tell us the number of ECTS you need for your degree program and whether it is graded.

Erasmus and Exchange Students

Please contact the instructors to make an arrangement. Tell us the number of ECTS you need for your university/degree and whether it is graded.

DEADLINE FOR ALL ESSAYS: 31.07.2025