MIND AND ARTIFICIAL INTELLIGENCE

PHIL.48J | SYLLABUS

Summer 2026 | Boğaziçi University

I. CONTACT INFORMATION

Instructor: Aran Arslan
E-mail: ararslan@gmail.com
Class Hours: Mond-Wed-Fri
Slots: 5 and 6
Office: TBA
Office Hours: By appointment via e-mail.

II. COURSE DESCRIPTION

What defines the mind? What is it to have a mind? How do thoughts and consciousness connect to the brain and body? What is the nature of intelligence? What is mental representation and content? These core questions in the philosophy of mind have regained prominence with the rise of modern Machine-Learning and AI systems, which raise the possibility of artificial agents replicating—or even replacing—human thought and behavior.

This course explores these debates by examining fundamental issues in the philosophy of mind—including the nature of thought, consciousness, and representation—and their relevance to current AI technologies like Large Language Models (LLMs). The philosophical focus includes mind-brain identity, functionalism, the hard problem of consciousness, qualia, content externalism, and mental representation. We will use AI as a case study to refine and deepen our understanding of the mind, and we will explore Mechanistic Interpretability (MI), which reverse-engineers neural networks into human-understandable explanations, and investigate conditions for Artificial General Intelligence and digital consciousness. Students will be encouraged to critically engage with these questions and form their own perspectives.

This is an advanced seminar typically for 3rd and 4th-year students. Although the course is accessible to non-philosophy majors, it demands sustained engagement with dense philosophical argumentation and careful reading of primary sources. This course is not recommended for students who need a passing grade in order to graduate. I strongly do not recommend you to take this course if you are not able to pay the required attention to it.

III. LEARNING OBJECTIVES

- **Grasp of Central Issues:** Students will gain a solid understanding of key debates in the philosophy of mind and AI, including the nature of thought, consciousness, and intelligence, and how these concepts apply to modern technologies like LLMs.
- **Critical Engagement:** Students will critically analyze major philosophical perspectives—such as mind-brain identity, functionalism, and externalism—and evaluate how AI challenges or reinforces traditional views of the mind.
- **Independent Thinking:** Through in-class discussions, students will develop and defend their own reasoned positions on both classic philosophical problems and contemporary AI-driven dilemmas.
- **Understanding AI Systems:** Students will understand the architecture of neural networks, transformers, and LLMs at a conceptual level through required video materials, enabling informed philosophical evaluation of these systems.

IV. SCHEDULE OF READINGS AND TOPICS

The following is the schedule of required readings and topics. Students are expected to complete all required readings and watch all required videos ***before*** the corresponding lecture. Readings will be shared at least one week before the corresponding lecture via email. **The required videos are not optional supplements**: technical content from them will appear in the short-answer sections of both the midterm and the final exam.

Main textbook: *David J. Chalmers (ed.) (2021), Philosophy of Mind: Classical and Contemporary Readings* (Second Edition), New York: Oxford University Press.

Wk	Topic	Required Reading(s)	Required Video(s)	Exam
1	**Introduction: The Mind-Body Problem**	**Descartes, René**, Meditations on First Philosophy (II and VI), in Chalmers (ed.)	*But what is a neural network? (3Blue1Brown)* https://youtu.be/aircAruvnKk	
2	**Identity Theory and Its Critics**	**Lewis, David (1966)** 'An Argument for the Identity Theory', Journal of Philosophy 63(1), pp. 17–25. **Kripke, Saul (1980)** Naming and Necessity (excerpts on the modal argument against type-type identity theory), in Chalmers (ed.).	*Gradient descent, how neural networks learn (3Blue1Brown)* https://youtu.be/IHZwWFHWa-w *Backpropagation, intuitively (3Blue1Brown)* https://youtu.be/Ilg3gGewQ5U	
3	**Functionalism**	**Putnam, Hilary (1973)** 'The Nature of Mental States', in Chalmers (ed.). **Block, Ned (1978)** 'Troubles with Functionalism' (Excerpt), in Chalmers (ed.).	*Backpropagation calculus (3Blue1Brown)* https://youtu.be/tIeHLnjs5U8 *Large Language Models explained briefly* https://youtu.be/LPZh9BOjkQs	
4	**Consciousness and Qualia**	**Nagel, Thomas (1974)** 'What Is It Like to Be a Bat?'. Philosophical Review 83: 435–50, reprinted in Chalmers (ed.). **Jackson, Frank (1982)** 'Epiphenomenal Qualia'. The Philosophical Quarterly 32: 127–36, reprinted in Chalmers (ed.).	*Transformers, the tech behind LLMs* https://youtu.be/wjZofJX0v4M *Attention in transformers, step-by-step* https://youtu.be/eMlx5fFNoYc	**Midterm**
5	**State and Content Externalism**	**Putnam, Hilary (1975)** 'The Meaning of Meaning' (focus sections: §I–§IV [Twin Earth and the rejection of methodological solipsism] and §VI [the division of linguistic labour]). Philosophical Papers, Vol. II: Mind, Language, and Reality. Cambridge University Press. **Burge, Tyler (1979)** 'Individualism and the Mental' (sections I–III only), in Chalmers (ed.).	*How might LLMs store facts* https://youtu.be/9-Jl0dxWQs8 *But how do AI images and videos actually work?* https://youtu.be/iv-5mZ_9CPY	
6	**The Digital Mind**	**Turing, Alan M. (1950)** 'Computing Machinery and Intelligence'. Mind 59: 433–60. **Searle, John (1983)** 'Can Computers Think?', in Minds, Brains, and Science, pp. 28–41. Harvard University Press, reprinted in Chalmers (ed.).	*The Dark Matter of AI [Mechanistic Interpretability]* https://youtu.be/UGO_Ehywuxc *Mechanistic Interpretability: A Whirlwind Tour* https://youtu.be/veT2VI4vHyU	
7	**AI, Interpretability, and Consciousness**	**Bereska, L. & Gavves, E. (2024)** 'Mechanistic Interpretability for AI Safety – A Review'. arXiv preprint arXiv:2404.14082. (Required: §1 Introduction, §2 Core Concepts, §3 taxonomy of methods, and §6 Discussion. Remaining technical sections are reference material.) **Chalmers, D. J. (2025)** 'Propositional Interpretability in Artificial Intelligence'. arXiv. https://doi.org/10.48550/arXiv.2501.15740.	*Could a Large Language Model be Conscious?* https://youtu.be/bskf9jyxmMs *Reading an AI's Mind with Sparse Autoencoders* https://youtu.be/krINuMZhJmU	
8	**Final Exam Week**	*In-class final examination (handwritten, closed-book). Covers all course material (Weeks 1–7).*	—	**Final**

Optional Readings by Week

Week 1 (Introduction: The Mind-Body Problem):

- **Kim, Jaegwon** (2022) 'The Mind–Body Problem after Fifty Years'. In A. O'Hear (ed.), The Philosophy of Mind, pp. 3–33. Cambridge University Press;
- **Fodor, Jerry A**. (1981). 'The Mind–Body Problem'. Scientific American 244: 114–25.

Week 2 (Identity Theory and Its Critics):

- **Smart, J. J. C**. (2022). 'The Mind/Brain Identity Theory'. Stanford Encyclopedia of Philosophy (Winter 2022 Edition).

Week 3 (Functionalism):

- **Kim, Jaegwon** (1992). 'Multiple Realization and the Metaphysics of Reduction'. Philosophy and Phenomenological Research 52: 1–26;
- **Davidson, Donald** (1970). 'Mental Events', in Chalmers (ed.).

Week 4 (Consciousness and Qualia):

- **Block, Ned** (1995). 'Concepts of Consciousness', in Chalmers (ed.);
- **Lewis, David** (1988). 'What Experience Teaches';
- **Jackson, Frank** (1986). 'What Mary Didn't Know'. Journal of Philosophy 83: 291–5.

Week 5 (Intentionality):

- **Brentano, Franz** (1995). 'The Distinction Between Mental and Physical Phenomena', in Chalmers (ed.).

Week 6 (The Digital Mind):

- **Berkeley, Istvan** S. N. (2019). 'The Curious Case of Connectionism'. Open Philosophy 2(1): 190–205;
- **Buckner, C.** (2019). 'Deep Learning: A Philosophical Introduction'. Philosophy Compass 14(e12625).

Week 7 (AI, Interpretability, and Consciousness):

- **Frankish, Keith** (2024). 'What are Large Language Models Doing?'. In A. Strasser (ed.), Anna's AI Anthology, pp. 55–78. Xenomoi.
- **Harding, Jacqueline**. (2023). Operationalising representation in natural language processing.

V. COURSE REQUIREMENTS AND GRADING

Component	Weight	Description
Attendance	**20%**	Tracked via in-class sign-in. **Missing more than 3 sessions results in automatic failure**.
Active Participation	**15%**	**Graded qualitatively per session (0/1/2 scale)**. Students must speak substantively, engage with peers, and reference the readings.
In-Class Midterm Exam	**25%**	End of Week 4. **Handwritten, closed-book. Covers Weeks 1– 4.**
In-Class Final Exam	**40%**	Separate exam week. **Handwritten, closed-book. Covers all course material (Weeks 1–7).**

Examination Format

Both the midterm and final examinations are handwritten, closed-book, and conducted in class. Each exam consists of two parts:

- **Part I – Short-Answer / Identification Questions:** Students will be asked to define key technical terms and explain their role in the relevant philosophical argument (e.g., "Define 'rigid designator' and explain its role in Kripke's argument against the identity theory").
- **Part II – Essay Question:** Three possible essay questions will be distributed to students in advance. One of these questions will appear on the exam. Students must write a sustained philosophical essay in response. The essay must reconstruct the relevant argument(s), engage with objections, and demonstrate independent critical thinking. The Part II essay question on the final examination will require substantive engagement with at least one reading from Weeks 5–7.

This examination format is designed to ensure that all assessed work reflects the student's own understanding and reasoning. ***No take-home written assignments will be given****.*

Attendance Policy

Students are expected to attend all sessions and arrive prepared to engage in discussion based on the assigned readings. Attendance is tracked via in-class sign-in.

Important: Missing more than 3 sessions (out of 21 total sessions) will result in automatic failure of the course. **No exceptions or grade negotiations will be permitted**.

Active Participation

Participation is graded qualitatively each session on a 0/1/2 scale:

- **0:** No substantive contribution.
- **1:** At least one substantive contribution (e.g., a question or comment that engages with the reading material).
- **2:** Multiple substantive contributions, including engagement with another student's point and direct reference to the assigned reading.

VI. COURSE POLICIES

Academic Integrity and AI Policy

Plagiarism—presenting someone else's thoughts as if they are your own—is a serious academic offense. This course has been specifically designed so that all assessed work is completed in class under examination conditions. The use of generative AI tools (such as ChatGPT, Claude, Gemini, or similar systems) to prepare for exams is permitted as a study aid, but all examination responses must reflect the student's own understanding and be written entirely by the student during the examination. Any evidence of academic dishonesty will result in automatic failure of the course.

Classroom Conduct

Cell phones must be switched off during class. Using laptops for purposes that are not class- related is prohibited.

VII. ADDITIONAL RESOURCES

For guidance on writing philosophical essays, students are encouraged to consult:

Jim Pryor, "Guidelines on Writing a Philosophy Paper":
http://www.jimpryor.net/teaching/guidelines/writing.html

Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/