Philosophy 283
Robot, Alien, and AI Consciousness
Course Syllabus
Winter 2025

Professor Eric Schwitzgebel
Office: 3208 HMNSS
Email: eschwitz@ucr.edu
Meetings: Fridays 2:00p-4:50p; remotely on Zoom at ucr.zoom.us/my/eschwitz

Topic
We will attempt to assess under what conditions we would be warranted in thinking that a robot, AI system, or naturally-evolved space alien would, or would not, be conscious. Readings will mostly be philosophy but will also include selections in science fiction, Artificial Intelligence research, and cosmology.

Graduate students from any department are welcome, and very advanced undergraduates with permission.

As an experiment, I have opened this class to philosophy PhD students and postdocs across the world. In a selection process based on c.v., writing sample, and statement of interest, I admitted five remote students, plus one independent researcher. To keep everyone on equal footing, the class will be conducted entirely remotely on Zoom.

Course Assignments
QCOs: Every participant in the course must prepare two questions/comments/objections (QCOs) on the reading every week except the first week. I may ask you to share your QCOs orally during seminar meetings. Each question/comment/objection should be one substantial paragraph long, so that a complete QCO is two paragraphs, about 300-500 words total. QCOs will be graded S/NC only. QCOs must be ready by the starting time of the seminar. Late QCOs are permissible only in the case of unforeseen emergency.

You must submit at least 6 pairs of satisfactory QCOs to pass the course. You will be given one free mulligan, so the expectation is that you will submit 8 satisfactory pairs. If you miss two QCOs (i.e., turn in only 7), your course grade will be lowered by 1/3 of a grade (e.g., A- to B+). If you miss three, your course grade will be lowered by 2/3 of a grade. If you miss more than three, the goddess Justicia will strike you down with her cold blind blade.

Auditors are also expected to contribute QCOs.

Final Papers: If you are taking the course for a letter grade, you must submit one final paper, due Thursday of finals week, connected with the themes and readings of the course. The paper should be 3500-5000 words. (Longer is acceptable with permission.) Graduate students from departments other than philosophy are welcome to submit a research project in the style of their home discipline. For example, a psychology student could submit a literature review paper or a detailed proposal for a series of experiments, or a science fiction student could submit a literary analysis or a creative work of science fiction.

Auditors and students taking the class S/NC are welcome to, but not required to, submit a final paper if they have a project related to the class on which they would like my feedback.

Readings

Availability: 32 of the 33 readings are available in a shared folder on Dropbox at: https://www.dropbox.com/scl/fo/qrdchc017w6yl6dl41vx8/AGnBRvHbFYEjruUvON2ZXUk?rlkey=ujv3fd18cjvmyeaxhpljdgs5q&st=3pgzg4jg&dl=0

Advice for the readings: The reading load is relatively high and some readings will be in fields or subfields (especially cosmology) where you likely won't know all the jargon and presuppositions. Do not exhaust yourself by trying to understand every word of every reading. Instead, read at variable speed: sometimes just skimming quickly for main points, at other times much more slowly where there's an argumentative turn crucial to an issue of interest to you. Different students will value different parts.

I recommend that you skim all the optional readings to see if you want to engage with them further. Some weeks, you might prefer to emphasize one or more of the optional readings over one or more of the required readings.

The one reading not in Dropbox is a novel by Greg Egan, *Diaspora*, to be discussed in Week 10. Purchase *Diaspora* well in advance and try to start reading it by the middle of the term so that you have time to complete it by Week 10.

Schedule

Jan 10 Introduction

Jan 17 Some classics of 20th century functionalism:

1. Putnam 1965: simple functionalism and the mind as computer
2. Lewis 1980: functionalism with natural kinds / identity theory, plus Martians
3. Levin 2003/2023: overview of functionalist positions

Jan 24 Tests of consciousness and organizational invariance

4. Turing 1950: locus classicus of the Turing test
5. Chalmers 1996: a slippery slope argument for functionalism
6. Schneider 2019: two potentially practical tests of AI consciousness
7. Optional: Udell & Schwitzgebel 2021: critique of Schneider (and Chalmers)

Jan 31 Critiques of functionalism:

8. Searle 1980: locus classicus of the Chinese room argument (skim or optional: critics' responses and Searle's replies)
9. Block 2002/2007: doubts about consciousness in functionally isomorphic robots
10. Optional: Baker 2016: SF story about biologicism gone wrong
11. Optional: Block 1978/2002: objections to functionalism, including Chinese nation
12. Optional: Cole 2004/2024: overview of Chinese room argument and objections

Feb 7 Large Language Models:

13. Chalmers 2023: LLMs might soon be conscious
14. Bender and Koller 2020: LLMs don't understand language, the octopus
15. Birhane and McGann 2024: LLMs don't understand language, embodied cognition
16. Optional: Schwitzgebel forthcoming: SF story with robot consciousness as an open question

Feb 14 Recent functionalist approaches:

17. Butlin et al. 2023: computational functionalist "indicators" of consciousness
18. Dehaene, Lau, and Kouider 2017: global workspace and higher-order consciousness

Feb 21 Biologicism:

19. Godfrey-Smith 2016: consciousness depends on low-level metabolic processes
20. Cao 2022: multiple realizability is less plausible than it seems
21. Dung 2024: consciousness depends on coarse-grained functional organization

Feb 28 Alien psychology:

22. Kershenbaum 2021: alien sociality and communication
23. Shostak 2010: aliens are likely to be AI
24. Bostrom 2014: what superintelligences are likely to want
25. Döbler and Raab 2021: "exopsychology" without anthropocentrism
26. Optional: Frank, Grinspoon, and Walker 2022: planet-scale intelligence
27. Optional: Schwitzgebel 2019: psychological features of a durable intelligence

Mar 7 Copernican Arguments:

28. Schwitzgebel and Pober 2024: a Copernican argument for alien consciousness, a mimicry argument against some robot consciousness
29. Gott 1993: the Doomsday Argument
30. Bostrom 2002: Anthropic principles
31. Optional: Lee 2019: alien quasi-consciousness

Mar 14 Diverse forms of technological and biological life:

32. Egan 1997: Forms of life in a post-scarcity technological world
33. Chiang 1998/2002: atemporal alien language and mind

Mar 20 Final paper due

Bibliography of Readings

1. Putnam, Hilary (1965). Psychological predicates. In W.H. Capitan & D.D. Merrill, eds., *Art, mind, and religion*. University of Pittsburgh Press / C. Tinling.
2. Lewis, David K. (1980). Mad pain and Martian pain. In N. Block, ed., *Readings in philosophy of psychology*. Harvard University Press.
3. Levin, Janet (2003/2023). Functionalism. Stanford Encyclopedia of Philosophy (Summer 2023 edition). URL: https://plato.stanford.edu/entries/functionalism
4. Turing, Alan M. (1950). Computing machinery and intelligence. *Mind, 59,* 433-460.
5. Chalmers, David J. (1996). *The conscious mind.* Oxford University Press. Pp. 247-275, 386-388.
6. Schneider, Susan (2019). *Artificial you.* Princeton University Press. Pp. 46-65, 159-160.
7. Udell, David B., and Eric Schwitzgebel (2021). Susan Schneider's proposed tests for AI

consciousness: Promising but flawed. *Journal of Consciousness Studies, 28* (5-6), 121-144.

8. Searle, John R., and critics (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 1,* 417-457.
9. Block, Ned (2002/2007). The harder problem of consciousness. In N. Block, *Consciousness, function, and representation.* MIT Press.
10. Baker, David John (2016). The hunter captain. *Escape Pod #526.* URL: https://escapepod.org/2016/03/29/ep526-the-hunter-captain
11. Block, Ned (1978/2007). Troubles with functionalism. In N. Block, *Consciousness, function, and representation.* MIT Press.
12. Cole, David (2004/2024). The Chinese Room argument. *Stanford Encyclopedia of Philosophy* (Winter 2024 edition). URL: https://plato.stanford.edu/entries/chinese-room
13. Chalmers, David J. (2023). Could a Large Language Model be conscious? Boston Review (Aug. 9). URL: https://www.bostonreview.net/articles/could-a-large-language-model-be-conscious
14. Bender, Emily M., and Alexander Koller (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 5185-5198.
15. Birhane, Abeba, and Marek McGann (2024). Large models of what? Mistaking engineering achievements for human linguistic agency. *Language Sciences, 106,* #101672.
16. Schwitzgebel, Eric (forthcoming). Guiding star of mall patroller 4u-012. *Fusion Fragment, #24.*
17. Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, Rufin VanRullen (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *ArXiv* 2308.08708. URL: https://arxiv.org/abs/2308.08708
18. Dehaene, Stanislas, Hakwan Lau, and Sid Kouider (2017). What is consciousness, and could machines have it? *Science*, *358* (6362), 486-492.
19. Godfrey-Smith, Peter (2016). Mind, matter, and metabolism. *Journal of Philosophy, 113,* 481-506.
20. Cao, Rosa (2022). Multiple realizability and the spirit of functionalism. *Synthese, 200,* #506.
21. Dung, Leonard (2024). Consciousness without biology: An argument from anticipating scientific progress. Manuscript in draft.
22. Kershenbaum, Arik (2021). *The zoologist's guide to the galaxy*. Penguin. Pp. 168-223.
23. Shostak, Seth (2010). What ET will look like and why we should care. *Acta Astronautica, 67,* 1025-1029.
24. Bostrom, Nick (2014). *Superintelligence.* Oxford University Press. Pp. 105-114.
25. Döbler, Niklas Alexander, and Marius Raab (2021). Thinking ET: A discussion of exopsychology. *Acta Astronautica*, *189,* 699-711.
26. Frank, Adam, David Grinspoon, and Sara Walker (2022). Intelligence as a planetary scale process. *International Journal of Astrobiology, 21* (2), 47-61.
27. Schwitzgebel, Eric (2019). A theory of jerks and other philosophical misadventures. MIT Press. Pp. 181-187.
28. Schwitzgebel, Eric and Jeremy Pober (2024). The Copernican Argument for alien consciousness; the Mimicry Argument against robot consciousness. Manuscript in draft.
29. Gott, J. Richard, III (1993). Implications of the Copernican principle for our future

prospects. *Nature*, *363*, 315-319.

30. Bostrom, Nick (2002). *Anthropic bias.* Routledge. Pp. 43-58.
31. Lee, Geoffrey (2019). Alien subjectivity and the importance of consciousness. In A. Pautz and D. Stoljar, eds., *Blockheads!* MIT Press.
32. Egan, Greg (1997). *Diaspora.* Millennium. [**NOTE this is not in Dropbox**]
33. Chiang, Ted (1998/2002). Story of your life. In T. Chiang, *Stories of your life and others.* Penguin.