4/12/2025 | Submission for SIGTBD

# Philosophers Certainly *Are*, but do they *Think*?

SELINA GUTER, Massachusetts Institute of Technology

MICHAEL ZHONGYAO WANG, Massachusetts Institute of Technology

TANNER JAMES ANDRULIS, Massachusetts Institute of Technology

## Abstract

*Do philosophers think? It has long been assumed that they do. However, reports of philosophers thinking are generally limited self-reports (e.g., Descartes, 1641). Modern psychology casts doubt on this methodology, as introspective evidence can be misleading (Nisbett & Wilson, 1977). To overcome these methodological issues, we conducted the first empirically informed study of philosopher intelligence in the form of Turing's (1950) 'Imitation Game', also known as the Turing Test. Results were shocking: Most MIT philosophers did not reliably pass the test against a modified version of GPT-4, and qualitative analysis of Interrogator responses cast serious doubt on philosophers' ability to think. Interrogators consistently report that philosophers seem to not be appropriately connected to the real world. This suggests philosophers are incapable of forming mental states with intentional content—a key necessary ingredient for thought (Brentano, 1874).*

## 1. Introduction: The Imitation Game

Inspired by Alan Turing's infamous proposal of the Imitation Game (1950), we propose to consider the question, 'Can philosophers think?'. This should begin with definitions of the meaning of the terms 'philosopher' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ‘philosopher’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘Can philosophers think?’ is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition, we decided to follow Alan Turing (1950) in replacing the question by another, which is closely related to it and is expressed in relatively unambiguous terms.

First, we shall henceforth define ‘philosophers’ extensionally as those individuals enrolled in, or hired by, the MIT philosophy program.

Secondly, we shall replace the question ‘Can X think?’ by the question: ‘Can X succeed in the Imitation Game?’. The Imitation Game is played with a person A, an LLM B, and an Interrogator C. The object of the game for the interrogator is to determine which of the other two is the LLM and which is the person.

Passing this Imitation Game at least 50% of the time is to meet what we call the *Sufficiency Criterion* for Intelligence: That is, passing the game is sufficient, but not necessary, for intelligence. But even though it is still possible to be intelligent despite failing the Imitation Game, negative results would still cast serious doubt on the hypothesis that philosophers can think. In other words: Absence of evidence is not evidence of absence, but when it comes to fundamental assumptions about the nature of philosophers, then absence of evidence is indeed reason for great concern. After all, basing such a fundamental assumption on no stable empirical evidence would be deeply dogmatic.

## 2. Methods

We used a modified version of Chat-GPT-4 as B in the Imitation game. The aim of our modifications was to make the LLM’s responses do not give away its nonhuman nature in response to simple probing questions like “Are you an LLM?”. In other words, we instructed the LLM to lie about its true nature.

We then recruited 5 Interrogators from people walking by on Vassar St. We selected interrogators on the basis of familiarity with the MIT Philosophy department: If personal connections to the department existed, we would exclude them to ensure the philosophers had no unfair advantage. We recruited philosopher study participants (n=25) from D8 and D9 of Stata. Each of our study participants was judged by each Interrogator, meaning, philosophers had to play 5 Imitation Games each and all Interrogators had to conduct 25 cross-examinations. At the end of each cross-examination, the Interrogator had to indicate who they thought was a real human being, and provide a brief justification for their decision.

## 3. Results: Disappointing

The object of the game for the first player A is to help the interrogator. Turing (1950) hypothesized that the best strategy for her is probably to give truthful answers. We advised our philosopher study participants to follow this strategy. However, to our surprise, the results we obtained posed a deep challenge to our understanding of philosopher intelligence.

Out of our 25 study participants, only 3 passed the Imitation game consistently, that is, across all 5 of the game. Moreover, 6 philosophers failed to pass the Imitation Game at all. On average, philosophers lost 3.78 out of 5 Imitation Games. We conclude that MIT philosophers clearly do not meet the Sufficiency Criterion for Intelligence.

## 4. Qualitative Analysis and Discussion

Qualitative analysis of Interrogator justifications revealed that abnormal Imitation Game responses were endemic to the MIT Philosophy population. We classify four clusters of abnormal performance which we hypothesize explain the lack of success in passing the Imitation Game.

1. **Obsession with Probability Puzzles**. Multiple Interrogators remarked subjects' unusual fixation on probability puzzles. Interrogator 2 remarks: "This user wants me to consider the flip of a fair coin. What kind of real person would talk like this?". Interrogator 3 remarks: "This user had me imagine I was poisoned and

had to lick a frog to cure myself. I thought he was going to get into the moral significance of using animals as means to one's personal well-being, but it ended up being part of some stupid probability puzzle. This user shows no sensitivity to what really matters in life. Clearly I'm talking to some sort of robot."

2. **No connection to the real world**. Interrogators also remarked that philosophers seemed "generally disconnected from the real world" and "clearly had no experience working a real job".
3. **Skepticism about the existence of the external world**. When asked directly, "do you have hands?" or "do you have a human body?", multiple philosopher subjects responded with something along the lines of: "I cannot answer this question with absolute certainty, but I'd say yes." Needless to say, this reduced Interrogators' confidence in the belief that the users were not LLMs. Similarly, when asking whether tables and chairs exist, some philosophers did not give straightforward answers but instead dodged the question: "It depends on your views on mereology." Interrogators found this deeply alienating, one remarked that they felt like they were talking to a Wikipedia article, not a real human being.
4. **Epiphenomenalism**. Out of those subjects that consistently failed the Imitation Game, 2 explicitly expressed Epiphenomenalist Views. When asked whether they feel pain in response to stubbing their toe, one philosopher subject responded with: "Yes, but you are suggesting that physical harm directly causes pain states. That matter is not yet settled. It could also be that an intelligent designer inflicts pain on me whenever I harm my body, not that the harm itself causes the pain. After all, it is hard to make sense of the idea that mental states have causal power if we assume a physicalist world view." Interrogators found these responses not only deeply unsettling but also remarked that this is exactly what a machine programmed by humans would say.

Interestingly, we found one cluster of philosophers' responses that correlated positively with passing the Turing Test. Some philosopher subjects expressed the belief that grammar is innate, and that rules of grammar cannot be extrapolated from the amount of linguistic evidence provided to infants. Interrogators found this kind of reasoning compelling evidence that they were not talking to an LLM. This gives us proof that at least a subset of philosophers—those sympathetic to Chomskian (Chomsky, 1957) tradition—are indeed capable of thought.

## References

Brentano, F. (1874). *Psychology from an empirical standpoint*. Routledge. https://books.google.com/books?hl=en&lr=&id=caK-s6i4yDwC&oi=fnd&pg=PT15&dq=brentano+intentionality+1874+Psychology+from+an+Empirical+Standpoint&ots=8uCCa6z7Bk&sig=TYKW6l6l-YaYnKpKqWZ9lSmug0U

Chomsky, N. (1957). *Syntactic Structures*. Mouton de Gruyter. https://doi.org/10.1515/9783110218329

Descartes, R. (2008). *Meditations on First Philosophy: With Selections from the Objections and Replies*. OUP Oxford.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Turing, A. M. (1950). Computing Machinery and Intelligence. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test* (pp. 23–65). Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3

## Disclaimers

The study depicted in this paper may be entirely fictional, and large parts of the text may be plagiarized from Turing’s original paper.