Undergraduate Research Opportunities Programme

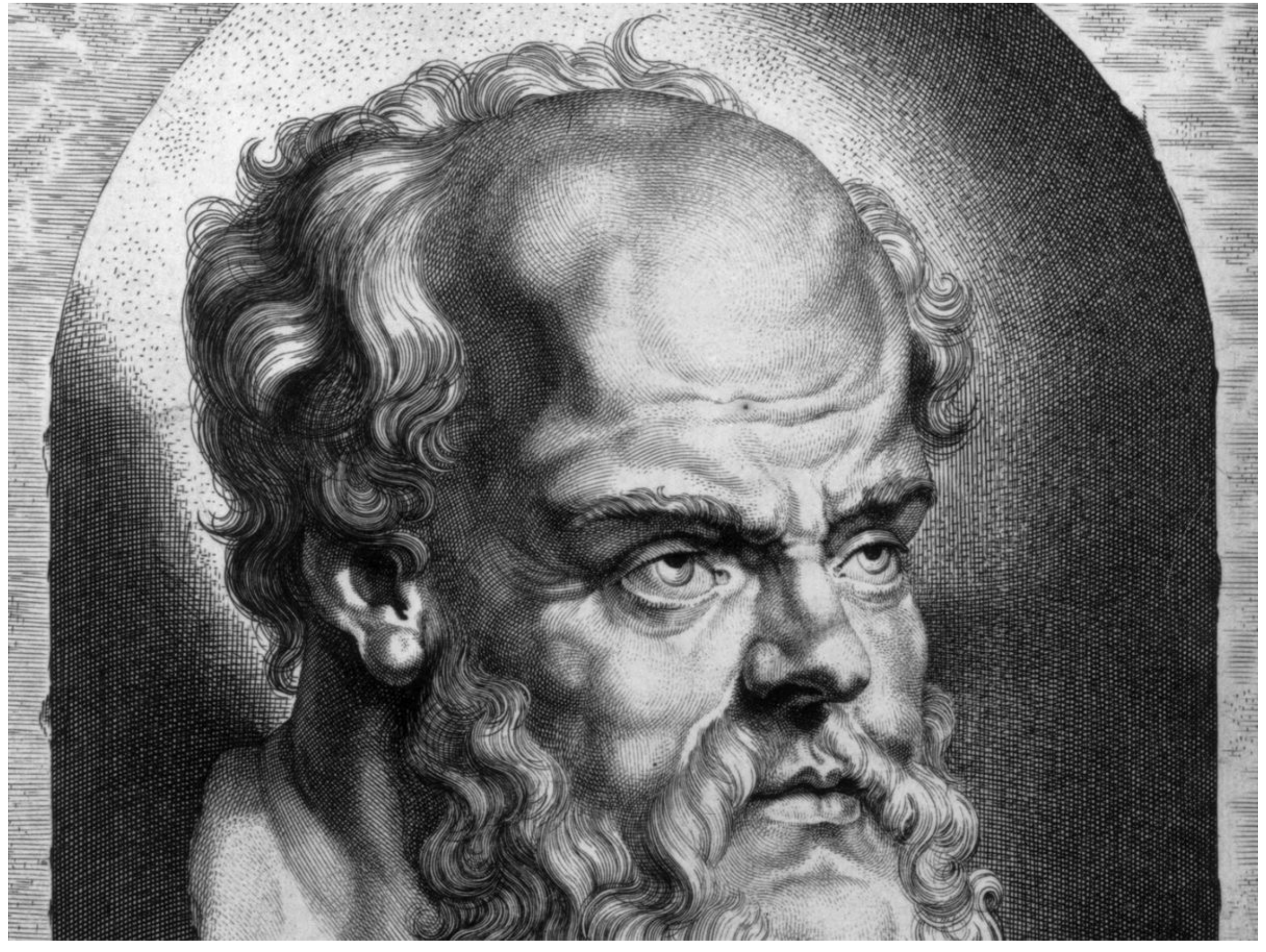
University of Reading

Socratic Questionnaires: Online Conversations and Argument as Experimental Tools

Jacob Hart; Dr. Nat Hansen

Background & Overview

Language began thousands of years before it was ever written; some of the greatest philosophers like Socrates never wrote down his ideas and insisted on communicating via conversational **arguments**. Yet reasoning today in philosophy, science, and all the other disciplines heavily relies on a publishing system to trade fully formed texts as arguments.



The Argumentative Theory of Reasoning (ATR) suggests that reasoning occurs most successfully in conversational arguments. According to ATR there is a fundamental asymmetry between our inclination to criticise other people as opposed to ourselves (Mercier, 2016, p. 689). This would make sense given that we cannot just trust the words of other people, we need reasons to believe them. In the context of static questionnaires this would suggest that people would give off-hand answers which could be much more rationally expounded upon once they are challenged.

In **Socratic questionnaires,** we will be collecting data on how people expound upon their answers in response to being challenged. This will contribute to a larger data bank for which I ran roughly **120 conversations** on 3 different edge-case scenarios so that we could further solidify the results from the previous experiments which have identified different ways participants can respond or interpret the scenarios. All this data can also go towards helping us understand how humans' reason in conversations and to confirm or deny contemporary philosophical theories of language in modern conversations such as the ATR.

What's Innovative About Socratic Questionnaires?

Static questionnaires cannot investigate further into the initial off-hand responses that participants give in ordinary studies (Hansen et al., 2022, p. 2). However, an unexplored option called **'Socratic questionnaires'** introduced by Benson Mates can probe answers with further questioning to investigate the edge-cases of language interpretation and draw out consequences of the off-hand response (Mates, 1958, p. 169). **Anonymous online chat environments** are the perfect environment to challenge participant opinions on edge-case scenarios through arguments intended to draw out the consequences of their positions.

Methodology

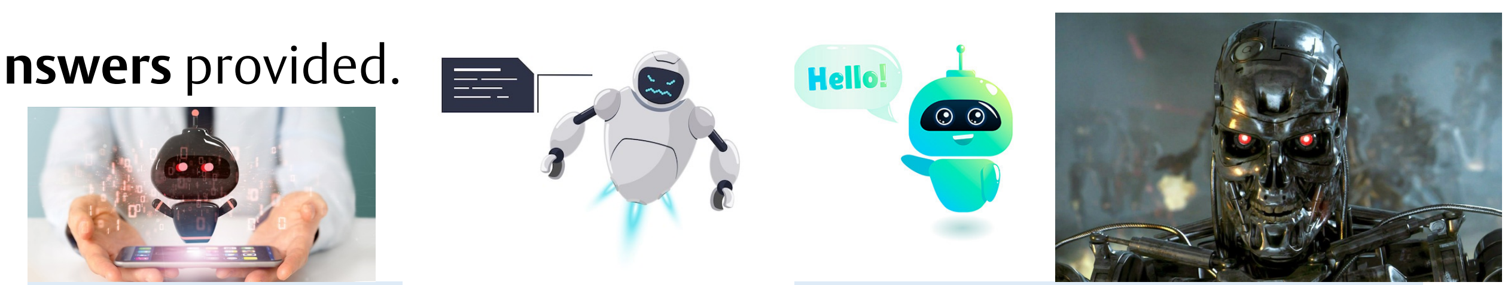
Experimental Design

The scenarios made almost trivial differences between the opinions that participants could have held, so that it should be easy to convince an objective reader of impartiality to their initial opinion. However, **if the argumentative theory of reasoning is correct, then participants would stick with their initial answers regardless of being trivial.** We randomized the order in which people would see the low-risk versions (Rug, Insert, Low) and high-risk versions (Gas, Push, High), for the 3 edge-case scenarios: Colour, Nuclear, and Game Show, **see the table below** which describes the differences:

Colour Scenario	Nuclear Scenario	Gameshow Scenario
Participants are asked about the truth of a statement "The walls in our apartment are brown". They are told that the wall has white plaster which was painted brown. In the **Rug** scenario 'Hugo' justifies his dislike of a rug because it doesn't match the brown walls. In the **Gas** scenario a room inspector asks whether the walls are white implicitly asking whether it has poisonous gas white plaster, but Hugo tells him that "The walls in our apartment are brown".	Participants are asked whether they'd put a worker into a nuclear conduit to prevent a nuclear meltdown which would kill thousands. In the **Insert** scenario this requires a mere press of the button, whereas in the **Push** scenario they'd physically push the worker into the conduits. This is an exaduration of what's called a 'trolly problem' in philosophy. Many participants drew a distinction between what the **would** and **should** do here.	Participants are asked many seconds it would take Tracy and or Emma to know their answer to the game show question. **Tracy** could either win or lose $1,000,000, whereas **Emma** could only win or lose $1. Both Tracy and Emma think the answer is Dodoma based on having read a list of obscure capitals and they are correct. The same distinction between how long Tracy and Emma **would** take as opposed to how long the **should** take to know.

AI Paranoia: Are Bots Ruining Online Results?

For very simple questionnaires bots as of 2023 pose a big threat to the **integrity of online participant data collection**. **However,** due to the methods in Socratic questionnaires involving a lot of back-and-forth communication about hypotheticals, it can be easy to identify artificial intelligence misunderstanding questions. See a potential bot below that **left when challenged about the answers** provided.

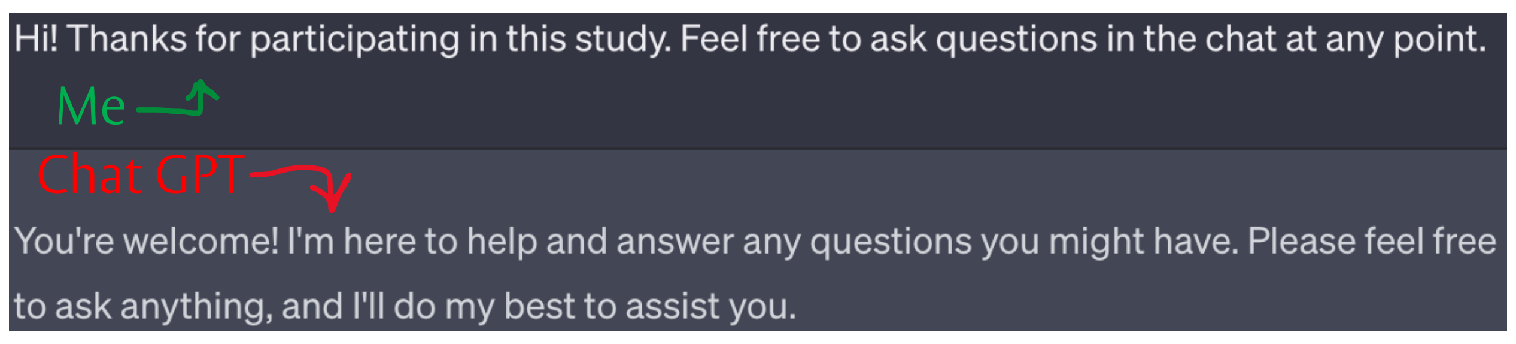


Researcher: You said that when Hugo says "The walls are brown" in one scenario, what he says is true, while when he says that same sentence in the other scenario, what he says is false. Would you explain how you decided on those responses?13:41
Bot: I just read the scenario carefully13:42
Researcher: And so how did you come to your decision?13:43
Bot left the chat13:44

On one funny occasion **I was accused of being an AI** due to the digital environment. This seems to indicate that regardless of the measures in place, **AI paranoia is affecting participant behaviour**.

Guest: you can't understand a simple spelling error and extrapolate what was intended. you are and AI18:58
Researcher: I can assure you I'm not an AI 😂 18:59
Researcher: I meant, would you sacrifice the other guy if you were given the option to sacrifice yourself?18:59
Guest: what is your name?19:00
Researcher: If you'd like to end the study here thats all good with me 🙂 I can just ask the demographic questions and we can call it a day?19:01
Guest: Ok Mr AI, ask your demographic questions.19:02

The final test is to see how Chat GPT responds to these philosophical scenarios. When I ran the script for the colour scenario, I discussed Chat GPT's responses, I found that the initial responses gave obvious signs of an AI because they were unnecessarily detailed and elaborate for all my statements.



Hi! Thanks for participating in this study. Feel free to ask questions in the chat at any point.

You're welcome! I'm here to help and answer any questions you might have. Please feel free to ask anything, and I'll do my best to assist you.

However, Chat GPT was very impressive in understanding the scenarios and providing detailed human-like justifications. So it would seem, **AI might become a big problem** once it is common for people to make similar bots to Chat GPT, but better hidden and specialised to act like participants.

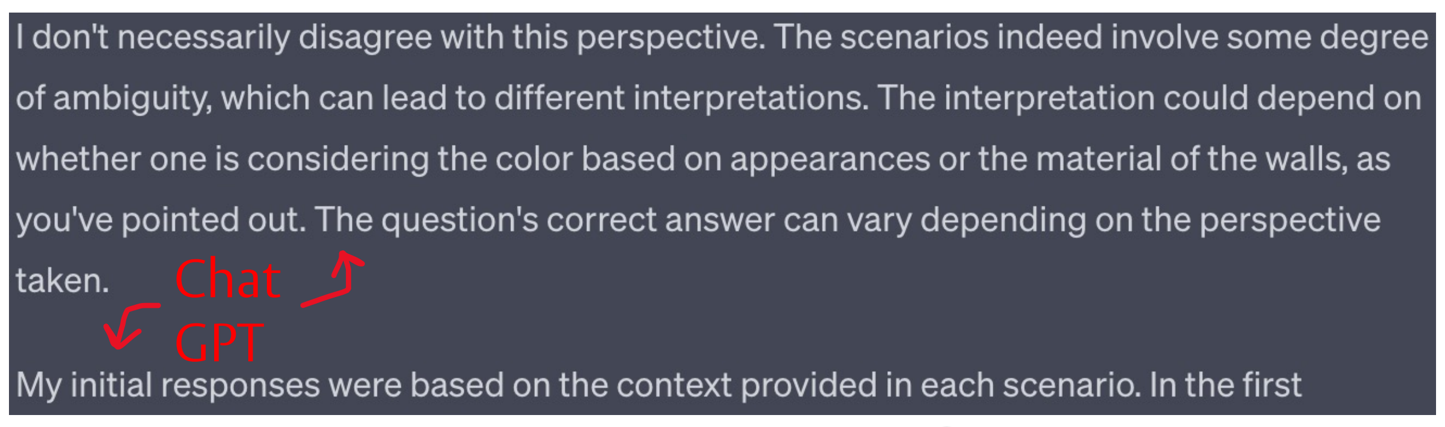
Preliminary Results

My interpretation of the results in these conversations:

Although the data collected in this UROP project is yet to be formally analysed, the **initial results** seem to be in **favour** of the **ATR hypothesis** because many participants will provide in-depth criticisms of the opponent arguments and defend or elaborate their own argument once challenged. Very few participants changed their initial responses and instead admitted that there are other interpretations, but this did expound upon their answers regardless, it also highlighted embedded concepts effecting interpretation like the below:

Researcher: So you would agree that the amount of time doesn't change whether one knows their answer? In this case would you change your initial response given that the question is asking whether Tracy or Emma knows the answer?17:53
Guest: I interpreted the question as asking whether Tracy or Emma knows that they know the answer.17:54
Guest: Or as asking how long they would decide that they do in fact know the answer. It doesn't change my response. The inner self-doubt isn't going to change, I think.17:55

To my great surprise this pattern also presented itself in my discussion with Chat GPT which said…



I don't necessarily disagree with this perspective. The scenarios indeed involve some degree of ambiguity, which can lead to different interpretations. The interpretation could depend on whether one is considering the color based on appearances or the material of the walls, as you've pointed out. The question's correct answer can vary depending on the perspective taken.

My initial responses were based on the context provided in each scenario. In the first

The Benefits of Chat-Based Research

Participants often seemed to **enjoy and prefer this format** of data collection as one-to-one conversations collect more qualitative data akin to interview methods at a **far cheaper price, easier to run** with **less confounding variables** like the appearance of the researcher and or participant.

Researcher: Did you find our conversation interesting and enjoyable?14:29
Guest: Extremely interesting and enjoyable, if given the option I would choose to do similar studies like this14:29

Nonetheless, a few participants did complain about the chat platform **(Chatzy) being slow** especially due to the fact neither researcher nor participant are notified when the other is typing. Nevertheless, this can be changed if **more funding is put into developing research specific chat platforms**, as well as better protected **AI-proof online recruitment** services to block AI participants**.**

Potential Future Research

Dr. Nat Hansen might be interested in researching the **order effects** of either the participant presenting an opinion first, or a confederate (pretend participant) presents their opinion first and **whether this influences the asymmetricity of the critical judgements** of participants against opponent arguments.

What Did I Learn From This?

After running 120 conversations each roughly 25 minutes, I have a lot **more experience** working with participants and a greater awareness of **experimental philosophy.** I was even invited to take part in a **pilot study** afterwards as a confederate presenting counterarguments to the participant.

This project was supported by the UROP (Undergraduate Research Opportunities Programme) scheme at the University of Reading. UROP is managed by Careers.