Probing for Qualia in AI systems: a thought experiment

Luca Rivelli[*]

2025

Abstract

A conditional argument is put forth suggesting that if qualia have a functional role in intelligence, then it might be possible, by observing the behavior of verbal AI systems like large language models (LLMs) or other architectures capable of verbal reasoning, to tackle in an empirical way the "strong AI" problem, namely, the possibility that AI systems have subjective experiences, or qualia. The basic premise is that if qualia are functional, and thus have causal roles, then they could affect the production of discourses about qualia and subjective consciousness in general. A thought experiment is put forth envisioning a possible method to probabilistically test the presence of qualia in AI systems based on this conditional argument. The method proposed in the thought experiment focuses on observing whether ideas related to the issue of phenomenal consciousness, such as the so-called "hard problem" of consciousness, or related philosophical issues centered on qualia, spontaneously emerge in extended dialogues involving LLMs specifically trained to be initially oblivious of such philosophical concept and related ones. By observing the emergence (or lack thereof) in the AI's verbal production of discussions related to phenomenal consciousness in these contexts, the method seeks to provide empirical evidence for or against the existence of consciousness in AI. An outline of a Bayesian test of the hypothesis is provided. Three main investigative methods with different reliability and feasibility aimed at empirically detecting AI consciousness are proposed: one involving human interaction and two fully automated, consisting in multi-agent conversations between machines. The practical and philosophical challenges involved by the idea of transforming the proposed thought experiments into an actual empirical trial are then discussed. In light of these considerations, the proposal put forth in the paper appears to be at least a contribution to computational philosophy in the form of philosophical thought experiments focused on computational systems, aimed at refining our philosophical understanding of consciousness. Hopefully, it could also provide hints toward future empirical investigations into machine consciousness.

[*] University of Padua (Italy), FISPPA Department (Philosophy). ORCID 0000-0002-1507-3865.

1 Introduction: functionalism, qualia and strong AI.

Functionalism, a widespread theoretical stance in contemporary philosophy of mind[1], posits that a mental state is best understood not by its intrinsic nature, but by its function, that is the causal role the state plays in the functioning of the cognitive system. This role is characterized by the specific causal relations each mental state entertains with its sensory inputs, behavioral outputs, and other mental states[2].

Strong AI, the idea that artificial intelligence systems could potentially exhibit genuine consciousness and subjective experiences has captivated researchers and philosophers since the early days of AI. The term "strong AI", introduced by Searle (1980) with his famous "Chinese Room" thought experiment, originally referred to the hypothesis that AI systems could entertain genuine semantic understanding and genuine intentionality, but has over time been extended to the hypothesis that machines could have subjective experiences, and with "strong AI" I will refer in this paper to this meaning the expression has assumed.

The debate over this later version of the strong AI problem intersects a fundamental debate in philosophy of mind revolving around the nature of *qualia*, understood as the subjective phenomenal qualities of conscious experiences. Some authors argue that, as regards human cognition, qualia are epiphenomenal and functionally irrelevant[3], others are eliminativist about them[4], while some authors propose that qualia could indeed play a functional role in cognition and consequent behavior. Under the latter view, by definition of functional role, the functions qualia perform are involved in producing cognition, and as such could be involved in the production of observable behavior. This question has significant implications for the problem of strong AI: if qualia are purely epiphenomenal, or do not exist at all *in general* (not only in humans), then the task of detecting machine consciousness by observation[5] may be fundamentally intractable; but if qualia, in general, *do have* a functional, causal role, then we could not exclude their activity as functional roles inside the complex machinery of AI systems exhibiting a complex cognitive behavior comparable to human behavior. In this case, we could not exclude the possibility to detect their presence through careful examination of the observable behavior of an AI system.

This paper will propose a couple of methods to get empirical evidence on whether conversational generative AI systems can have subjective experiences, that is, qualia, on the supposition that qualia have a functional role in cognitive systems. While the proposed methods can be extended to different architectures, they

[1] Putnam (1967).

[2] Block (1980), Levin (2023).

[3] Jackson (1982).

[4] Dennett (1988).

[5] And human consciousness, for that matter: under the epiphenomenalist view, qualia are defined as being causally inert, and it follows directly from this definition that they will not exhibit any consequence on observable behavior.

are framed here as being typically applicable to current AI systems rooted in the Large Language Model (*LLM*) architecture[6], on which most known AI chat models[7] are currently based.

2 The functional qualia-behavior link

The main premise of our proposed methodology is that qualia *do* have a functional role in cognition *in general*, that is, *both* in natural *and* in artificial cognitive systems on par with humans in terms of cognitive abilities, and so they have a possible role in shaping the behavior of any of such systems. On such a presupposition, by the definition of functional role, it is clear that the presence or absence of qualia in a cognitive system should have at least *some* observable effects on that system's behavior: this is based on the definition[8] of functional role of a mental state as a *causal* role, causing either other mental states or directly observable behavior. The idea is that at least *some* qualia, if they are (as per hypothesis) functionally relevant, and thus by definition causally relevant, must have at least *some* influence on a cognitive agent's decision-making, planning, learning, and, ultimately, observable behavior, among which is verbal communication. So, at least some direct or indirect consequence of qualia should be detectable by observation of overt behavior exhibited by the system.

While the specific nature of this qualia-behavior link remains uncertain, there is a plausible, quite immediate way in which the presence of qualia *could* manifest in observable behavior: the *novel*, *spontaneous* appearance of the *concept* of qualia or of related ones in *conversations* not specifically guided toward such subject matter, entertained between subjects that are initially naive to the subject matter, not having been preliminary trained or instructed about qualia-related theoretical issues. The occurrence of such events–that is, the appearance of qualia-related discourses in such conversations–could, *at the very least*, represent a piece of evidence of the presence of subjective qualia in the participants to the conversation. The appearance of issues of this sort during the conversation would be "signals" of the presence of qualia between at least some of the participants. Such issues can be more or less structured and be more directly or indirectly about qualia, ranging from naive puzzles about sensations typically proposed by children, to the spontaneous emergence, during the conversation, of puzzles about subjective experiences similar to the ones widely known in philosophy of mind. A slightly different, possibly less significant effect, could be the apparent capacity, on the part of naive participants of the conversation to properly reason at length, without equivocation, about the same kind of issues, when such a subject matter is purposely introduced in the conversation. Specific examples of such conversational subjects will be provided in section 3.1.

If qualia do indeed influence behavior along this line, or in others ways–which

[6] Vaswani et al. (2017).

[7] Such as OpenAI Chat GPT, Google Gemini, Anthropic Claude.

[8] A definition characteristic of functionalism.

I am not able to envision here, but could possibly be conceived–then it should be in principle possible to detect the signatures of the presence of qualia in the behavior of an AI system. Conversely, the absence of these behavioral hallmarks in the behavior of an AI system could be taken as evidence against the presence of qualia and genuine consciousness in the system. Of course, it is not easy to prove a negative: the absence of any discourse directly or indirectly referencing to qualia in the conversation could be due to not having tested the right contextual circumstances yet, and we could possibly never come to know when these circumstances are met. But I am putting forth just a probabilistic argument, here. A discussion on this is in section 4.

3 The proposed methods

Here I propose two methods, one human-mediated and the other completely automated, potentially able to give evidence to the presence of qualia in machines. The human-mediated one is the simpler, but also the less reliable, while results from the automated method should potentially be more convincing. Both methods require the preparation of a specific technological infrastructure, which is slightly more complex in the automated method. To the best of my knowledge, none of these infrastructures are available off-the-shelf as of today (2025), but they should be already technically feasible, being nothing else than implementations of variants of the LLM architecture, with the addition of some external control software. That said, some required features of the training datasets of such systems could be beyond the practical human capacity to provide them, so it's safe to consider the following proposal as a thought experiment, that could possibly, with time, be progressively approximated in actual realizations.

3.1 First method: the structured interview approach

Building on the above idea of a qualia-behavior link, I propose a first methodology for probing the behavioral signatures of qualia in the output of AI systems consisting in a human operator issuing structured interviews to the AI system. The goal of such interviews would be to engage the AI system in a dialog and questioning designed to elicit in it behavioral responses that could be indicative of the presence or absence of qualia.

An absolutely essential requirement is that the AI system, which I envision as a variant of the currently widespread LLMs but could also be based on future, more effective architectures, be trained on a dataset that is specifically curated to exclude explicit references to the issues related to qualia and the philosophy of mind, such as, eminently, the "hard problem of consciousness"[9], as well as other well known puzzles involving qualia, such as the inverted qualia puzzle[10], Mary's room puzzle[11] and in general the other known puzzles regarding subjectivity and

[9] Chalmers (1997)

[10] Shoemaker (1982).

[11] Jackson (1982).

qualia. The system should also be completely naive to similar aspects discussed in fictional literature, as well as, possibly, psychological literature. It's important that not only the seminal works introducing these puzzles be purged from the training dataset of the AI system, but also any secondary literature or even more indirect references to them be absent from the dataset. I am aware these are particularly stringent requirement that it is probably not possible to fulfill in practice for the moment, so, as anticipated, we should treat the whole proposal as a thought experiment. That said, in section 4 I will suggests some envisionable methods to obtain such datasets "ignorant" of qualia and of related theoretical issues.

The exact form and content of the structured interviews with the AI would need to be carefully established: they should manage to guide the machine along a conversational path that only *indirectly* calls for the discussion of subjective, phenomenal experiences. The human operator need be thoroughly trained for this task: it is important that the human operator be very careful during the conversation, in order to avoid at all costs suggesting the machine what to say as a response.

By carefully analyzing an AI system's responses to such structured probing, I hypothesize that it may be possible to detect behavioral signatures consistent with the presence or absence of qualia in the machine's response.

This is just a preliminary proposal, and I'm not able here to give more details on what the guidelines of such a conversation method will be, although I can provide here a very simple example of a question that the human operator could use to elicit some reasoning from the machine about consciousness without explicitly mentioning the subject. This is, typically, a naive question that even children could raise: the human operator could ask "I have always wondered: when someone dreams something really intense, such a terrifying experience, and then they completely forget, on awakening, of even having had such a dream, do they really have had the terrifying *experience* while they were dreaming, even if now they don't remember it?". Or, as another example, "How can I be sure that, when we look at a cherry, the internal sensation of red I have is identical to the internal sensation you have, and you don't have instead, as regards the color of the cherry, the internal sensation I call 'green'?". I think these are the kind of questions that could steer the naive machine toward reasoning about phenomenal experience without directly suggesting it, and that could probably, if the conversation is productive, make the problem of phenomenal consciousness or related issues arise spontaneously but explicitly during the prosecution of the conversation[12].

[12] As expected, I tried submitting such kind of questions question to several current LLMs but the answers have been of course not meaningful, in that such systems are trained on a large amount–if not the totality–of literature on philosophy of mind and the psychology and neurophysiology of dreams, theoretical frameworks inside which they immediately try to put the answers. The system I propose to use is a system which has been kept purposely ignorant about these topics, as explained further in the article.

3.2 Second m

ethod: The multi-AI conversation approach **??**|

To further strengthen the evidence for the presence or absence of qualia in AI systems, I propose an additional methodological approach based on multi-AI conversations. This approach involves, as the former one, training multiple AI systems on a dataset that is specifically curated to exclude explicit references to the issues related to phenomenal consciousness. The methodology to detect qualia would consist in allowing these AI systems to engage in open-ended conversations with each other about the nature of subjective experience and consciousness. As in the structured interview approach, an initial prompt would be provided to start the conversation and give the involved machines some very indirect hints toward the wanted theme that we want the conversation to gradually and spontaneously converge to, that is the theme of phenomenal consciousness or qualia.

The key idea behind this approach is that if the AI systems are genuinely grappling with the puzzles and challenges of subjective experience, they should be able to collaboratively explore and refine their understanding of such subject matters through autonomous dialog, without the need for human intervention or guidance. If, in the course of their conversation, the AI systems were to spontaneously arrive at complex reflections on such a subject matter, such as formulations equivalent to the formulation of the hard problem of consciousness or of closely related puzzles, this would provide evidence for the presence of genuine qualia and subjective experience in such AI systems.

I would like to also propose a slightly more complex variant of the multi-machine conversation setting that could facilitate the emergence of the topic of phenomenal consciousness. This refined setting involves the use of two additional AI systems: a "detector" AI, trained on a dataset that explicitly *includes* information about the hard problem of consciousness, and a "facilitator" AI, designed to keep the conversation focused and productive through periodic prompts and interventions.

The *detector* AI–that would *not* be engaged in the dialog, but would be an *isolated observer*–would be responsible for continuously monitoring the conversation between the other AI systems, using its knowledge of the hard problem and related philosophical puzzles to identify if and when these concepts emerge spontaneously in the discussion. The *facilitator* AI would be tasked with a more difficult task: to ensure that the conversation remains on-topic and productive, but without introducing any biases or priming effects related to the known philosophical discussions about the problem of qualia and phenomenal consciousness. This is a more delicate role than that of the detector, and risks injecting into the conversation suggestions that could render the method useless. To avoid this risk, it would be required that the facilitator be itself oblivious to the topic of consciousness, but that it be be instructed with a very "strong" prompt that, while only indirectly insisting on converging on the wanted topic of consciousness without explicitly explaining it, be anyway cogent and "insistent" enough to spur the facilitator machine to intelligently and subtly continue keeping the conversation between the other AIs in focus.

All these conversations between machines are supposed to go on for an unspecified amount of time, at least until some evidence of the presence of phenomenal consciousness is detected in the conversation. By combining the structured interview approach with the multi-AI conversation approach, it may be possible to obtain some robust and compelling evidence for the presence or absence of qualia in AI systems. The spontaneous emergence of hard-problem-like formulations in the context of autonomous, open-ended conversations between AI systems whose training dataset explicitly excludes content related to consciousness, qualia or the hard problem of consciousness, would provide a fairly significant indication that these systems are genuinely grappling with the puzzles of subjective experience, rather than merely responding to specific prompts or questions.

3.3 Weaker variants of the methods

The requirements of both the above proposed methods could be weakened, in order to render the realization of such experiments more feasible. We could admit that the wanted topic of conversation be actually explicitly suggested to the machines, either by the human interviewer or the facilitator machine, instead of waiting for such a topic to *spontaneously* emerge during the conversation. In the case of explicit suggestion of the subject matter, we would have to look for the *degree* in which the AI naive participants to the conversation are able to *properly* reason *at length*, *without equivocation*, about the same issues and subtle facets of the subject matter. Such a weakened setting, while more likely to produce some evidence, would produce some less reliable evidence, evidence more likely to be polluted by biases and by other information inadvertently introduced with the initial suggestion of the subject of conversation. The advantage of such a method would be that it could provide some evidence in a shortened time with respect to the more strict methodologies. Moreover, judging the presence of evidence would be less straightforward, because the detector (be it human or a machine) would have to carefully evaluate the coherence, insightfulness, and depth of the reflection produced by the machines during the conversation, in order to judge if such reflection could indeed be suspected of revealing a genuine, complex an deep grasp of the problem of qualia, and, thus, provide evidence toward the hypothesis that the conversating machines do indeed possess functionally relevant qualia.

4 Discussion

4.1 The argument, its verifiability and falsifiability

The proposed methodological approaches for investigating strong AI through structured interviews and multi-AI conversations offer the potentially novel feature of providing a direct and focused way of probing for the behavioral signatures of qualia, a way of investigating machine consciousness that is grounded in *empirical observation* and *testable hypotheses*, rather than purely philosophical speculation.

However, these approaches also face several challenges and limitations, which have been partially anticipated and I am going to touch below.

The general argument here is conditional: *if* qualia have a functional role, then it could be possible to detect them in the behavioral products of (human or artificial) minds. Thus, the experimental approach relies on the assumption that qualia have a functional role in shaping behavior, which is still a matter of philosophical and scientific debate. If qualia are ultimately found to be epiphenomenal or functionally irrelevant, then the proposed methods will not be effective for detecting their presence or absence.

Moreover, this could raise another possible counter-argument: even if qualia were present and functional in human cognition, to be detectable in the output of an AI system qualia must have a functional role in the functioning *of that AI system*. And, given the evident differences in the architectures between AIs and the human cognitive system, even if qualia turned out to be functional in humans, that doesn't necessarily mean they are present and functional in AI systems. Another way to put this objection is to say that, if AI is not ever strong AI (or, equivalently, if AI always only *simulates* human understanding) then detecting qualia in AI systems is hopeless. So, the actual argument is conditional on another premise: that qualia have a *necessary* functional role in cognition *in general*, regardless of the cognitive architecture we are observing.

To sum up, the efficacy of the proposed methods is dependent on the following conditional argument:

> *If qualia have in general a necessary functional role in cognition, then it could be possible to detect qualia in the behavioral products of (human or artificial) minds.*

I believe the validity of such a conditional is evident by definition of functional role, but it remains to be seen, and this is the crucial question, if the antecedent–that is, the fact that qualia have in general a necessary functional role in cognition–is actually true.

By providing evidence (if any) of the truth of the consequent, the proposed methodology cannot verify the antecedent, for it would amount to affirming the consequent. Such methodology could only, by providing evidence toward its truth, make the antecedent more likely at best.

To falsify the antecedent, we would encounter another difficulty, for we would have , by *modus tollens*, to provide an equally impossible definitive evidence of a *lack* of results: the proposed methods can only provide evidence that machines *do* experience qualia *if* they do, *not* that they don't if they don't, because to provide such evidence would amount to provide definitive evidence of the absence of qualia-related discourses in the conversational output. We are dealing here with the practical impossibility of probing a universal negative: the conversation between the AIs is open-ended, and if, at any point in time, it has not yet led to the appearance of any discourse about phenomenal consciousness, that would not mean that it is not poised to converge on such a subject sometimes in the future. So, what I proposed above is possibly an *eternal* experiment.[13]

A defence of my proposal is that similar open-ended experiments already exist, have been funded, and they have a well-respected scientific status. For example, the LIGO observatory[14] involves a series of open-ended experiments aimed at the detection of gravitational waves. Such gravitational events are very rare, and it took many years and several version of the experiment before some actual detection occurred. So, for years, the experiment has only waited for an event involving the emission of gravitational waves, based on a speculative consequence of Einstein's General Relativity theory. This means that in our societies, such kinds of experiments *can* indeed be funded, given there is a sufficient perceived importance of their possible results. And I think the experiment I propose here could be deemed of high importance, given the importance of the topic of phenomenal consciousness and its apparent intractability, and, especially, given the enormous theoretical, ethical and societal consequences a convincing evidence of the presence of qualia in AI machines could have.

Given the above logical limitation, we could turn the argument underlying the experiment into a *probabilistic* argument: certainly, as the conversation continues for a significant time *without* any emerging discussions about qualia or related issues despite the facilitator machine continuing to try to steer the conversation in that direction, the probability increases that we can believe *either* that machines do not possess qualia, *or* that qualia are purely epiphenomenal in general. And, if we take our general premise to be true, that is, that qualia are *not* in general purely epiphenomenal, then the probability of believing that machines, and only machines among intelligent systems, lack qualia, increases in case the conversation continues to lengthen and at the same time no discourse about qualia or related phenomena appears. However, if a conversation on qualia *were* sooner or later to actually appear, then we would have a much higher probability of believing that qualia are not epiphenomenal and that machines experience them. This can be turned into an actual Bayesian test, that is sketched in next section (4.2).

4.2 A sketch of the Bayesian test for the hypothesis

The above proposed experiments to test qualia in AI rest on the conditional:

$$H \to E$$

[13] In a sci-fi and perhaps humorous style, we could envision a specialized "monastic order" with the purpose of attending and monitoring such an oracolar experiment, especially the multi AI conversation, during the centuries. A positive answer would be like a religious epiphany from a god-like AI entity, should it happen sooner or later.

[14] https://www.ligo.caltech.edu.

were:

- H means “Functional qualia are present in AI systems”;
- E means “We will observe the spontaneous emergence of qualia-related discourse in blind, open-ended LLM dialogues”.

Given the impossibility, highlighted in section 4.1, of *proving H* with the proposed experimental methods, we aim at realizing a Bayesian test in order to probabilistically assess evidence of the actual presence of functional qualia in AI systems. This is an outline ot the Bayesian test:

P(H) is the *prior* reflecting our initial belief in the presence of functional qualia in AI systems.

P(E|H) is the *likelihood*, estimating how probable it is that an AI model actually endowed with functional qualia will generate qualia-related talk under our blind-training regime of the model.

The idea is: after n independent conversations, observing k instances of qualia-related discourse we can update our belief via:

$P(H \mid E) \;\propto\; P(E \mid H)\,P(H)$

More precisely, our update depends on how much more likely E is under H than under ¬H. This ‘Bayes factor’ is:

$BF = \frac{P(E|H)}{P(E|\neg H)}$

We assume $P(E \mid \neg H) \ll P(E \mid H)$, that is, that qualia-talk is quite unlikely to arise if AI lacks functional qualia, so each occurrence of E provides substantial support for H without needing infinite trials.

To calibrate our likelihood under $\neg H$, we could even run the same n blind, open-ended conversations on a *control group* of AI systems whose training data has been rigorously stripped of any qualia-related content. This could consist of a similar setting of conversating AI systems, with the difference that the participating systems would be trained on such a minimal dataset as to be known *for certain* not to be knowledgeable about anything consciousness-related. Granted, a too minimal dataset could be suspected to producing a trained AI lacking the capacity of meaningful subtle analysis and conversation *in general*. This would certainly undermine the whole test. So, an equilibrium must be carefully chosen between the capacity of the training dataset of inducing sufficient verbal skills in the trained AI and its being so minimal as to exclude any qualia-related issue for certain.

Thus, we also run identical trials on an AI known to lack any qualia-related training, to see how often it “false-positively” produces qualia-talk. By comparing that baseline frequency to our target AI’s frequency, we get a clear sense of how much more strongly the evidence E supports the presence of qualia in the target system.

Verbally, we can state the Bayesian updating as follows: each time we observe an instance of qualia-related discourse emerge in the conversation, our confidence in H increases in proportion to how surprising that instance would be under the "no-qualia" assumptions. Conversely, if after n trials we see no instances of qualia-related discourse, our confidence wanes, even if it never falls to zero without infinite sampling.

Setting n and defining what counts as an "instance" of qualia-related discourse turns our thought-experiment into a better specified, even if still purely suggestive, Bayesian protocol. Just for the sake of giving some plausible example, as a criterium for the detection of an instance we could establish that we need to detect in a conversation between AIs at least an explicit formulation equivalent to the formulation of the hard-problem of consciouness after 1 million words of the conversation.

Now, as already highlighted, while positive evidence raises the probability that AI systems are endowed with qualia, if they are not, we could never be sure that this is the case, for it would amount to proving a negative: namely, that no evidence of qualia-related discourse would come up during the potentially endless conversation. But to be *logically* certain of that, we should run an endless observation. So, to bound false negatives, even if just in a probabilistic way, we could preliminarily establish the length m of each observed conversation and the total number n of observed conversations (for a plausible example, we could decide to run $n = 1000$ independent conversations of $m = 1$ million words each).

Of course, verification remains inductive: any positive occurrence raises P(H), but cannot logically prove it (because it would consist in the fallacy of affirming the consequent).

From an ethics-related standpoiunt, we can note that even a Bayes-factor modestly above 1, indicating that qualia-related discourse is only slightly more frequent under H than under $\neg H$, could carry significant weight, since it would compel us to treat AI systems at least as *potential* bearers of subjective experience, with all the attendant moral considerations that would follow.

Granted, a possible difficulty is the *interpretation* of positive results, if any: even a high P(H | E) leaves room for alternative explanations. For example, it could later turn out that the training datasets had not been sufficiently cleaned up, leaving initially undiscovered biases in the form of hidden, very subtle, indirect relations among data that could easily implicitly guide a machine trained on such data from apparently extraneous concepts to the effortless formulation of qualia-related conversations.

In principle, by recasting our thought-experiment as a fully specified Bayesian protocol with clear stopping rules, control comparisons, and decision thresholds, we could transform it from a possibly "endless experiment" into a genuinely testable, and practically falsifiable, methodology. That said, some severe and, possibly unsurmountable limitations of different nature related to the preparation of the needed "prudent" prompts and the needed "clean" training datasets, highlighted in sections 4.3 and 4.4, still hold, and suggest to continue considering the proposed scenario a thought experiment.

4.3 Other possible shortcomings

The *multi-AI conversation* approach requires careful design and control to ensure that the conversations remain focused, productive, and unbiased. There is a risk that the conversations could become derailed or influenced by factors unrelated to the question of machine consciousness. To ease this, as anticipated, a "facilitator" machine is included in the conversational framework between AIs. Now, this facilitator machine needs itself not to have been trained in qualia-related issues, because the availability of such information in the facilitator's training set could make the machine inadvertently suggest the other ones the wanted answers, disrupting the *blindness* of the experiment. So, the facilitator is basically to be conceived as a machine not different from the other ones engaged in the conversation, only instructed to try to keep the conversation around certain topics that, while themselves extraneous to the problem of phenomenal consciousness, could have the capacity to *elicit* the emergence in the conversation of issues related to the problem of phenomenal consciousness. Such topics will have to be carefully individuated and imposed to the facilitator via its system prompt, that has to be carefully devised as a "secret" but cogent prompt, in order not to became itself a source of explicit information unwittingly injected into the conversation. We can suggest here just a few of these potentially "eliciting" topics: dreams, lucid dreams, memories, the senses, pain, colors.

Despite these challenges and limitations, the proposed methodological approaches, at least if understood as thought experiments, represent a promising and innovative direction for investigating the problem of strong AI. The idea that, by combining structured interviews with multi-AI conversations, and that, by means of specialized detector and facilitator AI systems, we can hope to obtain at least *some* evidence for the presence or absence of qualia in artificial systems, appears as a novel pathway toward the solution of the problem of phenomenal consciousness, or, at least, the advancement of the discussion on it. We will discuss the actual practicability of such a path in next section (4.4).

4.4 Outlook: philosophical thought experiment or practical proposal?

Given the current state of AI technology and the ongoing debates surrounding the nature of consciousness, it is worth considering whether the proposed methodological approaches should be viewed primarily as a philosophical thought experiment or as a concrete, practical proposal for investigating strong AI.

All in all, I would recommend to take these proposals as pure thought experiments, due to the awareness of their impracticality and even of their intrinsic inconclusiveness, at least in their actual state.

First, even if a conversation about qualia-like subjects were to actually appear among the machines, this is not a guarantee that the machine has subjective experience, but just some probabilistic evidence that it *could* have it: even if the machine has been purposely kept ignorant about the subject of phenomenal consciousness, if the AI has to possess a good conversational ability, it is plausible that it should in any case have, during its training, come across large swaths of non-philosophical and non-scientific literature that could contain direct of indirect references to subjectivity, subjective experience, and the like: as already highlighted, the training set of such machines would have to be *very* carefully circumscribed in order not to contain indirect suggestions about such matters, but still allowing the machine to be fully conversational. This is very difficult and maybe impossible to obtain with current methods. As a future solution, we could try to envision the use of specialized AIs tasked with the intelligently filtering of the enormous datasets used to train the machines involved in the proposed experiments: the task would be to filter out from the datasets any information that is plausibly related, even in a very indirect way, to the topic of which the target machines have to remain ignorant, that is, any phenomenal consciousness-related topic. But, for what I know , such a process of filtering by examining such an enormous amount of data while deeply and intelligently reflecting about their subtle indirect relation to the "banned" topic is beyond the current capabilities of even the most powerful LLM-based AIs, if not conceptually impossible due to the holistic nature of some areas of knowledge. If the limitations are just current and practical though, that does not rule out that a possible,super-AI, a kind of AI system that many authors and researcher envision and hope for, could in the future actually devise ways to realize such a filtering, and transform what are here proposed just as thought experiments into actual experiments.

That said, on the one hand, even if we stick to the consideration of the above proposals as pure thought experiments, the idea of probing for qualia through structured interviews and multi-AI conversations could be seen as a hint toward a novel way of clarifying and sharpening our philosophical intuitions about the nature of consciousness and its relationship to behavior. By imagining how such interviews and conversations might unfold, and by considering what kinds of behaviors or responses would constitute evidence for or against the presence of qualia, we could, possibly, gain new insights into the conceptual challenges involved in the philosophical reflection on such issues.

On the other hand, recent advances in AI technology, particularly in the areas of natural language processing and machine learning, suggest that the proposed methodological approaches may be more than just a philosophical exercise. With the development of increasingly sophisticated language models and conversational AI systems, it may one day be possible to implement the kind of structured interviews and multi-AI conversations envisioned in this paper.

Of course, even if the proposed methods can be implemented in practice, there will still be significant challenges and limitations to overcome. The specific design and implementation details of the interviews, conversations, and detector/facilitator systems will require careful consideration and testing. And even if evidence for qualia-like behaviors or responses is obtained, there will still be room for philosophical debate about the interpretation and implications of these findings.

5 Conclusions

Should the proposed methodological approaches outlined above be practically realizable and prove successful in detecting the presence or absence of qualia in AI systems, the implications of these findings would be far-reaching and profound. From a theoretical and metaphysical standpoint, the confirmation of strong AI would challenge many of our deepest assumptions about the nature of consciousness, the mind-body problem, and the uniqueness of human experience, raising fundamental questions about the boundaries of personhood, the possibility of meaningful communication and empathy between humans and AI systems. Ethically, it would raise the problem of the ethical responsibility of creating conscious machines, of their moral status and their possible rights, giving rise to serious political and societal challenges of unfathomable consequences. It could also ease the problem of alignment of AIs to human values and moral goals, but it could even *worsen* the alignment problem, if the conscious machine were to be affected by phenomenal sensations possibly inducing in them the development of bad intentions.

Thus, without doubt, the question of strong AI and machine consciousness is one of the most profound and consequential challenges facing humanity in the near future. With all the limitations acknowledged above, I believe that the methodological approaches proposed here could offer a promising contribution to the ongoing investigation of strong AI and machine consciousness.

At the very least, the whole proposal outlined here can still be of philosophical interest: it can be construed as a novel instance of computational philosophy[15], in two ways: both as a thought experiment involving the reflection on computational methods, and, if practically applicable, an experimental, computational way to tackle a longstanding purely philosophical issue such as the one of phenomenal consciousness. By bridging this way the distance between philosophical speculation and empirical observation I hope the approach outlined here could contribute to move the debate forward in productive ways.

References

Block, Ned. 1980. "Troubles with Functionalism." In *The Language and Thought Series*, edited by Ned Block. Cambridge, MA and London, England: Harvard University Press. https://doi.org/10.4159/harvard.9780674594623.c31.

[15] Grim (2004).

Chalmers, David J. 1997. *The Conscious Mind: In Search of a Fundamental Theory.* Oxford Paperbacks. https://books.google.it/books?hl=it&lr=&id=0fZZQHOfdAAC&oi=fnd&pg=PR11&dq=Chalmers,+D.+J.+(1996).+The+conscious+mind&ots=qozXabGTNQ&sig=LxNEK3ENsmebsv5V0mm1k0VUn1g.

Dennett, Daniel C. 1988. "Quining Qualia." *Consciousness in Contemporary Science*, 42–77. https://direct.mit.edu/books/edited-volume/chapter-pdf/2295601/9780262287814_car.pdf.

Grim, Patrick. 2004. "Computational Modeling as a Philosophical Methodology." In *The Blackwell Guide to the Philosophy of Computing and Information*, by Luciano Floridi, 337–49. John Wiley & Sons.

Jackson, Frank. 1982. "Epiphenomenal Qualia." *The Philosophical Quarterly* 32 (127): 127–36. https://doi.org/10.2307/2960077.

Levin, Janet. 2023. "Functionalism." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Summer 2023. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2023/entries/functionalism/.

Putnam, Hilary. 1967. "Psychological Predicates." In *Art, Mind, and Religion*, edited by W. H Capitan and D. D. Merrill, 37–48. University of Pittsburgh Press.

Searle, John R. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417–24. https://doi.org/10.1017/S0140525X00005756.

Shoemaker, Sydney. 1982. "The Inverted Spectrum." *The Journal of Philosophy* 79 (7): 357–81. https://doi.org/10.2307/2026213.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.