Supplementary Discussion for

Stefano Palminteri, Charley M Wu, Beyond computational equivalence: the behavioral inference principle for machine consciousness, *Neuroscience of Consciousness*, Volume 2026, Issue 1, 2026, niag002, https://doi.org/10.1093/nc/niag002

Orthogonal axes of the attribution problem

We define the problem of attribution as the question of whether or not a system implements a specific cognitive process. In order to determine the appropriate method, we propose to decompose this attribution problem into two different axes (**Figure S1**). The first axis concerns the *generality* of the attribution (from *momentary* instances to *ontological* statements). The second concerns the *method* used (from *folk* psychological heuristics to the *scientific* method).

Generality

Within the generality axis, we identify two levels. The first level can be called "*Momentary attribution*", which is concerned with answering the question of whether or not a given cognitive process is currently ongoing in a specific individual at a specific moment. A typical question would be: *is this person conscious right now?* The second level can be called "*ontological attribution*", which is concerned with the broader question of whether a class or group of entities possesses the capacity to express a given cognitive process. A typical question would be: *are non-human animals conscious*? Or, *are large language models (LLMs) conscious?*

Momentary attribution

Momentary attributions are generally carried out for entities that we have no reasonable doubt they would be conscious (e.g., other humans), and are based primarily on inferences from their behavior. For example, if you bump into someone in a shopping mall and they react by telling you to watch where you are going, you would immediately infer that they are conscious and aware of your presence. By contrast, the total absence of any observable reaction would suggest you have bumped into a mannequin rather than another conscious individual.

This form of evidence-based momentary attribution can also be flexibly revised "on the fly." If you see a friend lying on the couch, snoring, you may initially infer they are currently sleeping and not conscious. However, when they suddenly "wake up" and reveal they were only pretending to sleep (to play a prank on you), you would then immediately revise your attribution. Yet, basing momentary attributions on behavioral inference does not imply that they are blind to priors. If you are greeted by a humanoid robot at the ATR laboratories in Japan, despite its impressive range of human-like behaviors, you might be far more wary to attribute consciousness to it.

Crucial to the question of considering computational equivalence as a necessary criterion for attribution (see **Main Text**), momentary attribution of consciousness has occurred routinely since the dawn of humanity in the absence of any knowledge about the computational mechanisms underlying consciousness. This is highly relevant for the debate about whether we should apply double standards when attributing consciousness to humans versus machines. Momentary attribution simply relies on the assumption that a particular class of entities (e.g., humans) generally possesses the capacity for consciousness, and then infers whether the behavior under observation is best explained by assuming that the cognitive process is *currently* expressed. These attributions are pragmatic: their value lies in whether they allow us to explain, predict, and interact effectively with the individual in question. In other words, momentary attributions (of consciousness, or of cognitive processes in general) are instances of inductive inference and, more specifically, *inference to the best explanation*.

Ontological attribution

Ontological attribution starts from the relatively uncontroversial claim that there exists at least one class of beings of which we are certain they are conscious: namely, humans. This attribution comes naturally from our first-hand phenomenal experience and from sound similarity-based heuristics (other humans are physically and behaviorally very similar to us).

The ontological attribution question is then extended to other species, and the inference typically proceeds using a mixture of strategies, depending on the degree of similarity between the reference species (humans) and the target species. Consider, for instance, the question of attributing consciousness to chimpanzees, which are physically and behaviorally similar to humans. Thus, their similarity would hardly require much additional evidence to attribute a wide range of conscious processes to them. In this case, attribution can be largely heuristic and may not require sophisticated testing, experimentation, or hypotheses about the computational structure of consciousness (e.g., *folk attribution*, see below).

By contrast, ontological attribution becomes much less straightforward for more distant species, such as octopi (a question that remains unsettled). Here, heuristics informed by physical and behavioral similarity are far less informative. Thus, attribution must be mediated by an additional inferential step: octopi's behavior must be interpreted in terms of putative computational processes, and these inferred processes must then be compared with our current hypotheses about the computational processes underlying consciousness in humans. Importantly, this comparison is never a strict or literal evaluation of computational equivalence. There exists no "code" from which we can directly read the functions mediating cognitive processes in humans and octopi, which are implemented in radically different physical substrates. Instead, in both cases, the cognitive processes are postulated based on their effectiveness in explaining and predicting a given set of behaviors: again, an inference to the best explanation.

Methodology*:*

Within the methodological axis, we also identify two levels that roughly correspond to two different levels of methodological sophistication involved by the attribution decision. The first level, "Folk attribution" is based on simple and automatic heuristics, while the second level, "Scientific attribution" requires more sophisticated forms of evidence and inference.

Folk attribution

The attribution of consciousness—whether at the level of a single subject in a particular situation (*momentary*) or at the level of a broad class of entities (*ontological*), can be achieved through different methods. The *folk attribution* process relies primarily on innate or learned heuristics that are deployed almost automatically. For instance, as an example of momentary attribution based on folk psychological inference, we attribute consciousness to an awake and speaking human being, because we know that spoken language does not occur in the absence of consciousness (or only in very rare cases). As an example of ontological folk attribution, one might consider (again) the case of chimpanzees, to whom consciousness is largely attributed on the basis of their important physical and behavioral similarity with humans (although scientific experimentation may be useful to further refine our understanding). A crucial point is that folk attribution does not require and is routinely performed without committing to an explicit hypothesis concerning the computational structure of the process under investigation.

Scientific attribution

At the other extreme of the methodological axis lies the scientific method of attributing consciousness. At this level, attribution does not proceed via heuristics but must rely on rigorous experimentation, confirmation or falsification through empirical data, and at least some degree of explicit scientific hypothesis or theory. Scientific attribution is also technically orthogonal to the level of generality. For example, scientific attribution can be deployed "locally" to detect consciousness in locked-in patients using EEG. In this case, the inferential process is clearly evidence-based, not heuristic, and it still relies on a relatively modest theoretical background (in this case that certain neural signatures are associated with conscious processes). Yet, scientific attribution can also occur at the Ontological level, for instance, when we investigate whether octopuses (or large language models) are conscious.

Conclusions

Thus, there is no universal recipe for attributing cognitive processes, both momentarily ("is this person conscious right now?") and ontologiocally ("do octopi have consciousness?"). It can proceed in a largely heuristic fashion (*folk attribution*), relatively independent of formal computational theories, as in the case of closely related primates. But as the target species diverges further from the reference case, heuristic strategies become less reliable, and the process must instead be mediated by the inference of plausible computational processes from observable behavior.

When it comes to artificial intelligence and LLMs, the situation is even worse. Structural similarity with humans is entirely absent. Behavioral similarity is restricted to a single domain, language, which, although mastered to unprecedented levels, remains only one aspect of the broad repertoire of behaviors associated with consciousness. Consequently, the prior probability of LLMs being conscious is very low, and the burden of behavioral evidence required to reasonably attribute consciousness to such systems is correspondingly very high. That said, there is no a priori reason to deny that at least some forms of consciousness may have emerged if, after careful experimental and inference, this provides the most plausible explanation of their behavior.

To conclude, despite their differences, the dimensions of generality and methodology share some important features. Although in principle orthogonal, most practical cases fall along the diagonal (Figure S1). Momentary attributions are often based on immediate behavioral evidence and do not require intermediate inferences about computational mechanisms. By contrast, Ontological attributions often demand more formal experimentation and computational theorizing. Yet both approaches combine priors (based on heuristics, such as physical or behavioral similarity) with empirical evidence (behavioral or neural), and both ultimately rely on an inductive process.

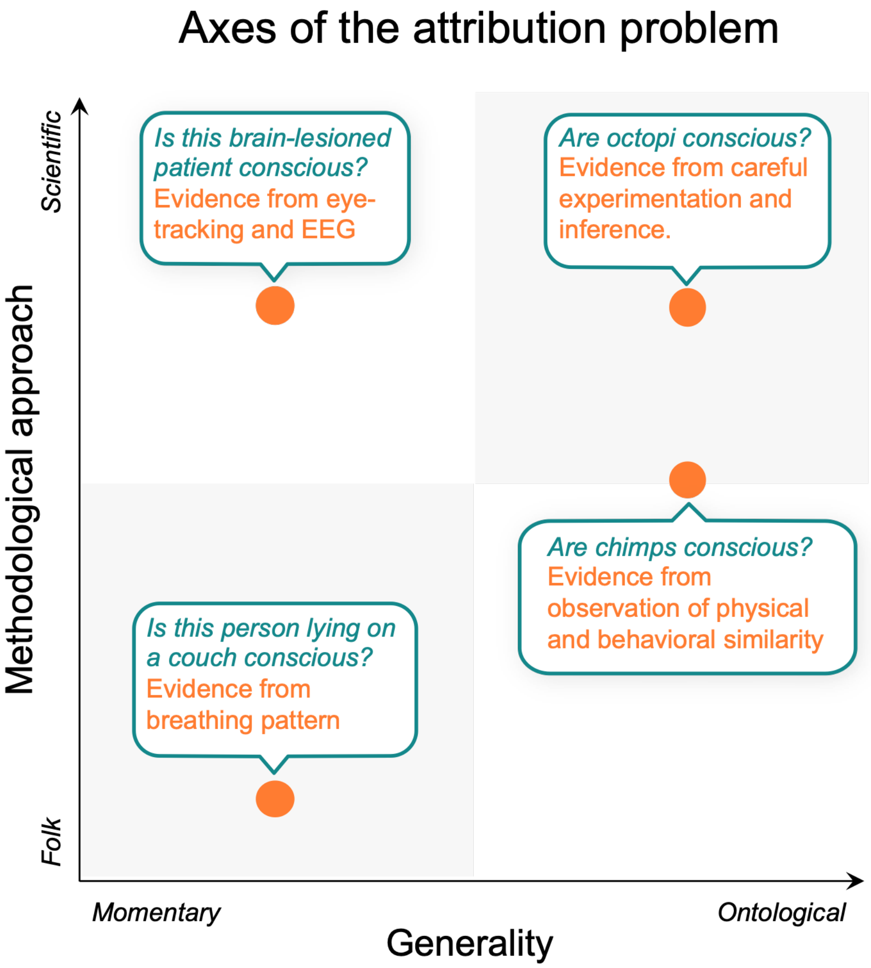
Axes of the attribution problem
Scientific
Methodological approach
Folk
Is this brain-lesioned patient conscious? Evidence from eye-tracking and EEG
Are octopi conscious? Evidence from careful experimentation and inference.
Are chimps conscious? Evidence from observation of physical and behavioral similarity
Is this person lying on a couch conscious? Evidence from breathing pattern
Momentary
Ontological
Generality

Figure S1: axes of the attribution problem. Visual representation of where different approaches to consciousness attribution map onto the axes of generality and methodology. Given the high physical and behavioral similarity between humans and chimps, the attribution of consciousness to this species can largely rely on folk assessment (such as those simple heuristics based on similarity). However, fine-grained questions concerning introspection and other language-based manifestations of consciousness still need to be addressed scientifically.