*** This is a preprint version of a paper forthcoming in Philosophy of Science***

Extrapolating other consciousnesses: the prospects and limits of analogical abduction

Dr. Niccolò Negro (corresponding author)
ORCID: 0000-0002-1561-799X
School of Psychological Sciences, Tel Aviv University
niccolonegro@tauex.tau.ac.il; niccolo.negro.research@gmail.com

Prof. Liad Mudrik
ORCID: 0000-0003-3564-6445
School of Psychological Sciences, Tel Aviv University
Sagol School of Neuroscience, Tel Aviv University
Canadian Institute for Advanced Research (CIFAR), Brain, Mind, and Consciousness Program, Toronto, ON, Canada
mudrikli@tauex.tau.ac.il

Advances in animal sentience research, neural organoids, and artificial intelligence reinforce the relevance of justifying attributions of consciousness to non-standard systems. Clarifying the argumentative structure behind these attributions is important for evaluating their validity. This paper addresses this issue, concluding that analogical abduction – a form of reasoning combining analogical and abductive elements – is the strongest method for extrapolating consciousness from humans to non-standard systems. We argue that the argument from analogy and inference to the best explanation, individually taken, do not meet the criteria for successful extrapolations, while analogical abduction offers a promising approach despite limitations in current consciousness science.

1. Introduction

Recent advancements in consciousness research, animal sentience research, neural organoids, and artificial intelligence have made the epistemological problem of other conscious minds – what, if at all, justifies the attribution of phenomenal properties to other entities – more relevant than ever. Since finding consciousness in such systems is likely to have significant ethical and societal implications (Shepherd 2018; Siewert 1998; Levy 2024; Birch 2024), this problem has become especially pressing. Accordingly, tests for consciousness are repeatedly being discussed and suggested (Bayne et al. 2024; Dung 2022; Andrews 2024; Kazazian, Edlow, and Owen 2024; Negro and Mudrik 2025; Schneider 2019), in an attempt to better understand the distribution of consciousness in non-trivial cases, both within the human population and outside of it. These populations include, among others, Artificial Intelligence (AI) systems, non-human animals, neural organoids, infants, and fetuses (see Table 1 for referrals to relevant literature on each population). Here, we refer to these populations as "non-standard systems" and frame the epistemological problem of other conscious minds in terms of *other consciousnesses,* asking how the current science of consciousness justifies generalizations about consciousness in these "non-standard systems"[1].

[1] We take to be neurotypical adult humans as examples of "standard systems", for which the presence of consciousness is not doubted. The primary generalization we are interested here is thus between these standard systems and non-standard systems, rather than between me and other people, as in more traditional discussions of the problem of other minds.

Table 1

Population	**Key papers**
Artificial Intelligence (AI) systems	Butlin et al. 2023; Chalmers 2023; Dehaene, Lau, and Kouider 2017; Dung 2023; Elamrani and Yampolskiy 2019; Hildt 2022; Schneider 2019; Sills et al. 2018.
Non-human animals	Andrews 2024; Barron and Klein 2016; Birch 2022; Birch, Schnell, and Clayton 2020; Carruthers 2019; Dung 2022; Halina, Harrison, and Klein 2022; Tye 2017; Veit 2022.
Neural organoids	Bayne, Seth, and Massimini 2020; Birch and Browning 2021; Croxford and Bayne 2024; Hameroff and Muotri 2020; Jeziorski et al. 2023; Lavazza and Massimini 2018; M. Owen et al. 2023.
Infants	Dehaene-Lambertz 2024; Passos-Ferreira 2023, 2024.
Fetuses	Bayne, Frohlich, Cusack, Moser, and Naci 2023; Ciaunica, Safron, and Delafield-Butt 2021; Frohlich et al. 2023; Moser et al. 2021.

Research on consciousness tests focuses mostly on the type of data (e.g., markers; Andrews 2024) that can serve as evidence for attributing consciousness to various target systems. Here, we address instead the complementary issue of defining the *reasoning* that underlies justified attributions of consciousness to different target systems, independently of the type of evidence one decides to exploit.

Thus, the driving question of this paper targets the logical structure of the reasoning we employ to address the epistemological problem of other consciousnesses: What is the strongest inferential machinery we could use to justify the attribution of conscious properties to non-standard systems? An agreement on the logical basis of our attribution practices is needed to clarify the argumentative structure that consciousness researchers ought to employ when concluding that a system has, or does not have, phenomenal properties. This is important both for assessing existing arguments for consciousness in non-standard systems, and for formulating future arguments of this sort.

Traditionally discussed within the more general problem of other minds, the epistemological problem of other consciousnesses has been approached through two different forms of reasoning: *analogical reasoning* and *reasoning from the inference to the best explanation* (IBE-based reasoning, for short). These are further rooted in two different inferential schemata: inductive inference for analogical reasoning, and abductive inference for IBE-based reasoning. In the philosophical literature, these types of reasonings have often been presented as *prima facie* competing and incompatible. For example, Hyslop (1995) champions analogical reasoning while exhibiting the flaws of IBE-based reasoning, while Pargetter (1984) does the opposite.

This attitude is partly mirrored in the current consciousness science literature (see Heyes 2008 for a related discussion focused on the science of animal consciousness): on the one hand, scholars discussing and developing different consciousness tests (e.g., the command-following test; see Owen et al. 2006 and Bayne et al. 2024 for discussion) extrapolate consciousness via analogical reasoning; on the other hand, Tye (2017) suggests an inferential strategy that can be seen as similar to the IBE-based reasoning, while Chalmers (1996) and Passos-Ferreira (2023) explicitly adopt it.

However, as suggested by Melnyk (1994), these two strategies are not logically incompatible, and indeed, in the current neuroscience of consciousness, many attributions of consciousness seem to incorporate aspects of both analogical reasoning and IBE-based one (e.g., Barron and Klein 2016). Birch (2022) puts it explicitly: "What we should do […] is build up a list of the behavioural, functional and anatomical similarities between humans and non-human animals, and use arguments from analogy *and* inferences to the best explanation to settle disputes about consciousness" (Birch 2022, 134; italics added).

Here, we further develop this approach, and provide a philosophical backbone to justify it, suggesting that the conjunction of analogical reasoning and IBE-based reasoning is the most promising approach when trying to determine which systems/organisms are conscious. We propose that the argument from analogy and the IBE-based argument are compatible and complementary, and that they can be fruitfully combined to deal with the epistemological problem of other consciousnesses. We do so by introducing analogical abductive arguments, and by showing that they can be used to overcome the problems that afflict analogical reasoning and IBE-based reasoning. We accordingly aim to provide a general structure for the 'inferential machinery' (i.e., argument) that can be used to address the epistemological

problem of other consciousnesses, independently of the specific fuel (i.e., test/evidence) one can put in that machine.

Hence, our project has both descriptive and normative components. The descriptive goal of the project is to substantiate analogical abduction as a way for capturing and systematizing a type of inferential practice that relies on both analogical and IBE-based strategies for extrapolating consciousness. We do so by introducing a novel argument schema that incorporates elements of both strategies. The normative aspect of our analysis adds a further layer: we argue that analogical abduction is the most compelling inferential strategy for dealing with the epistemological problem of other consciousnesses. Accordingly, we suggest that consciousness science would benefit from adopting this form of reasoning to systematically build and assess arguments for the attribution of consciousness to non-standard systems.

Moreover, we take the epistemological problem of other consciousnesses to be primarily concerned with the distribution of consciousness problem (i.e., is this system conscious or not?), rather than with the quality of consciousness problem (i.e., *how* is the system conscious/what is the system conscious of?) (Andrews 2024), so we will frame our discussion to address the distribution question. However, analogical abductive arguments can be leveraged to address the quality question as well.

In Section 2, we present the two traditional forms of reasoning identified in the philosophical literature to approach the epistemological problem of other consciousnesses, namely the argument from analogy and the IBE-based argument. In section 2, we introduce two challenges that any form of extrapolative reasoning must meet to be successful. In Section 3, we show that analogical abduction is a promising account to deal with the epistemological

problem of other consciousnesses. In Section 4, we consider some limitations of this account, and conclude that, although analogical abductive arguments cannot currently provide a definitive solution for the epistemological problem of other consciousness, they have the potential to do so, and therefore constitute our best available option.

2. The Epistemological Problem of Other Consciousnesses and the Two Traditional Forms of Reasoning to Approach it

The epistemological problem of other consciousnesses can be seen as an instance of a more general philosophical problem, the problem of *extrapolation* (Baetu 2024): how can one justifiably generalize from an epistemically privileged domain to a less epistemologically privileged domain (Steel 2007; Guala 2010; Thagard 1988)? Following the standard use in philosophy of science, we will refer to the epistemologically privileged domain as the 'SOURCE domain' (SOURCE), and to the domain of interest as the 'TARGET domain' (TARGET).

Depending on the scope of the extrapolative argument for other consciousnesses, SOURCE and TARGET can be identified in different ways. For the purposes of this paper, SOURCE will normally refer to the domain of neurotypical adult humans, from which the science of consciousness gathers most of its knowledge and upon which theories of consciousness are generally built and tested (Seth and Bayne 2022; Yaron et al. 2022; Mudrik et al. 2023; Mudrik et al. 2025) while TARGET will normally refer to non-standard systems in general.

Extrapolations in consciousness science would be fairly easy to justify if models of SOURCE-consciousness (i.e., human consciousness) were built in a context-independent way. That is, if the claims made about consciousness were evidently true irrespective of the characteristics of the SOURCE population. For example, theories of consciousness could be formulated in terms

of causal powers or capacities, which are by definition context-independent[2] (Hiddleston 2005; Cartwright 1994; Steel 2007, Ch. 5). If this were the case, theories of consciousness could explain consciousness by pointing at universal laws, and therefore, by employing explanatory constructs that are not dependent on the particular domain of applicability (in the same way as gravitational laws are supposed to apply to apples as well as to distant planets). This would make theories of consciousness conform with the requirement of universality (Kanai and Fujisawa 2024), rendering explanations in consciousness science closer to explanations based on universal laws as in physics. For example, the Integrated Information Theory (Albantakis et al. 2023; Tononi et al. 2016) aspires to provide such context-independent explanatory structure, given that it seeks to explain consciousness by relying on the notion of cause-effect powers of the physical (but see Merker, Williford, and Rudrauf 2021 for criticisms of its ability to do so; Lau and Michel 2019; Mediano et al. 2022). Nevertheless, the theory's axioms are based on phenomenological explorations of human experience, so the foundation of the theory might still be context-dependent, despite the proclaimed aspiration (see Bayne 2018).

Independently of how specific context-independent explanations of consciousness are constructed, the more general point is that it is questionable whether explanations in the biological sciences should indeed follow the same explanatory practices used in physics

[2] This is because capacities are supposed to be intrinsic features of an entity that are supposed to be stable across different background conditions: one could explain the combustion of wood by referring to the wood's capacity to burn. The power, or capacity, might not be manifested if the contextual conditions are not right (e.g., lack of oxygen), but it is supposed to exist nonetheless.

(Craver 2007, 2002), especially because many biological phenomena seem to be domain-dependent (e.g., an explanation of digestion in humans does not apply to cows, and theories of protein synthesis might not generalize to extraterrestrial life). It seems to be an open question, then, whether consciousness is the type of phenomenon that should be accounted for by universal generalizations, or whether, instead, its explanation should be domain-dependent.

This paper surveys some possibilities to attribute consciousness to other systems via extrapolative inferences, even if the explanation of consciousness indeed turns out to be context-sensitive and not universal. We will focus on the two *prima facie* different and alternative strategies suggested in the philosophical literature for formulating extrapolative inferences: IBE-based reasoning (Pargetter 1984) and analogical reasoning (Hyslop and Jackson 1972; Hyslop 1995); for a general introduction, see (Avramides 2000).

Both strategies seem well-suited for tackling the epistemological problem of other consciousnesses, because they build upon *ampliative* inferences, in which the conclusion conveys more information than the premises. We briefly present them below.

a. IBE-based reasoning

IBE-based reasoning exploits abductive inferences, namely inferences drawn in virtue of the explanatory power of the inferred hypothesis (Lipton 2004; Psillos 2002). The standard example is to infer, from the observation of wet streets, that it might have rained last night, since this conjecture is the best explanation of the evidence.

This argumentative strategy can be applied to the epistemological problem of other consciousnesses by noticing that some publicly observable properties[3] of a system of interest are best explained by the hypothesis that consciousness is required for their instantiation. The argument (adapted from Psillos 2002, 614) can be formalized as follows:

> P1. D is a collection of data about publicly observable properties of system S in TARGET.
>
> P2. The hypothesis H that S is conscious explains D (would, if true, explain D).
>
> P3. No other hypothesis explains D as well as H does.
>
> Therefore,
>
> C. H is probably true (i.e., S probably is conscious).

b. Analogical reasoning

Although there is much debate on how to properly characterize analogical arguments (for a comprehensive discussion, see Bartha 2019, 2010), a general enough form of analogical reasoning can be captured in the following way: we are justified in inferring that two systems are similar along certain unobserved dimensions if they are also similar with respect to some

[3] For the purposes of this paper, we lump together both neurobiological evidence and functional/behavioral evidence under the umbrella-term of "publicly observable properties". We remain neutral here on whether the best way to approach the epistemological problem for other consciousnesses is via the neurobiological route or the functional/behavioral one – see Block 2007 and Usher, Negro, Jacobson, and Tsuchiya 2023 for discussions.

observed dimensions, given prior knowledge that, in a given domain, the observed and unobserved dimensions of interest co-occur (Bartha 2019; Hesse 1965).

This reasoning can be applied to the case of other consciousnesses: given that I know that certain brain structures and activity, and/or certain functions and behaviors, reliably and systematically correlate with certain conscious properties in us (i.e., neurotypical adult humans), I can infer that similar conscious properties will be present in a system with brain structure and dynamics, and/or functions and behaviors, analogous to ours.

The analogical argument for other consciousnesses can be formalized as follows:

> P1. D is a collection of data showing that there is a systematic and reliable correlation between publicly observable properties and consciousness in the SOURCE domain (i.e., neurotypical adult humans)
>
> P2. D* is a collection of data about publicly observable properties of system S in TARGET
>
> P3. D* suggests that publicly observable properties of S are similar to those of SOURCE (i.e., neurotypical adult humans)
>
> Therefore,
>
> C. S probably is conscious.

3. How to Extrapolate Successfully

What does it take for an extrapolation to be successfully implemented? Following Steel (2007), we posit that any successful extrapolation must solve two problems: first, the *extrapolator's circle*: how to say something informative about the phenomenon in TARGET

given only partial knowledge of the target system and without assuming the presence of the phenomenon in TARGET. Second, the *problem of difference:* how to justify inferences about the phenomenon in TARGET given relevant dissimilarities between SOURCE and TARGET (this pair of problems was originally introduced by LaFollette and Shanks 1996).

We first examine how IBE-based reasoning might deal with these challenges. On the one hand, this strategy is not directly threatened by the problem of difference because it does not explicitly rely on similarities between SOURCE and TARGET. Moreover, it can solve the problem of difference by denying that differences between SOURCE and TARGET are explanatorily relevant for consciousness. This requires our best explanation of SOURCE-consciousness to successfully discriminate between properties (and their dimensions) that are relevant for consciousness, from properties (and their dimensions) that are irrelevant for consciousness. Arguably, this requirement is problematic given the current theoretical landscape, as it is questionable whether it holds for any of the presently available explanations/theories of consciousness. However, this is a flaw of current theories, not of the argumentative strategy itself, so we set it aside for now; let us assume that this problem can be solved by the IBE-based approach.

Even if so, we argue that this strategy fails to solve the problem of the extrapolator's circle. To explain why this is the case, we should first clarify exactly which cog in the IBE-based argument for other consciousnesses links what we know about SOURCE to what we say about TARGET.

This link is found in P2 in the above-mentioned schema for IBE-based arguments: 'The hypothesis H that S is conscious explains D (would, if true, explain D)'. Here, D refers to data about a system in TARGET, more precisely about publicly observable properties of the

system, but why are we justified in connecting such data to consciousness? In other words, why is H better than an alternative hypothesis (H*) that posits that those data can be explained by unconscious processes? If we want to select H over H* without appealing to similarities between SOURCE and TARGET (since that would push the argument towards an argument from analogy), then we need to *assume* that a well-established explanation of consciousness based on knowledge gathered in SOURCE is also applicable to the purported connection between publicly observable properties and consciousness in TARGET. But whether such explanatory connection is justified *in TARGET* is precisely what we need to establish, and therefore cannot be assumed.

To clarify this point: let us take SOURCE to denote neurotypical adult humans, and assume, for the sake of the argument, that we have a well-established theory built upon and tested on members of SOURCE. This theory can provide the means to determine if consciousness is indeed the best explanation for D. But since the theory was developed and tested on members of SOURCE, then it is *prima facie* a theory of *human*-consciousness (or of SOURCE-consciousness). The problem arises when we want to apply that theory to a non-standard system, which exhibits some interesting publicly observable properties, and argue that the best explanation for those properties is consciousness, based on the theory we have. This is problematic since those properties are explanatorily linked to consciousness *in the human case*: are we justified in considering the human-based theory as explanatorily powerful in the case of the non-standard (possibly non-human, too) system or not? (for a similar point, see Block 2002, Dung 2022, and Usher et al. 2023). This is precisely the epistemological problem of other consciousnesses, and assuming that we are in fact justified in drawing an explanatory connection between publicly observable data and consciousness *in TARGET*, as P2 in the IBE-based argument above implies, amounts to circular reasoning.

Thus, the IBE-based argument for other consciousnesses does not seem to have the resources, in itself, to avoid the extrapolator's circle. Again, this is the problem of explaining why we can gain knowledge about certain properties of TARGET given limited knowledge about TARGET, without assuming that those properties occur in TARGET to begin with.

Can analogical reasoning succeed where IBE-based reasoning fails? Analogical reasoning does not seem to be necessarily affected by the extrapolator's circle, because consciousness in TARGET is not assumed but rather *projected*; that is, rather than being inferred in virtue of an explanatory link that is assumed to be valid at the beginning of the investigation, it is instead inferred in virtue of some similarities between SOURCE and TARGET[4].

However, analogical reasoning fails to solve the problem of difference (i.e., the problem of explaining why certain unobserved similarities between SOURCE and TARGET should be present, given the relevant dissimilarities between the two domains). This is due to the inevitable presence of relevant differences between SOURCE and TARGET: how much difference can we accept without considering SOURCE and TARGET too distant for the analogy to hold? And what should our criteria for determining some threshold for answering this question be? According to Steel, "any adequate account of extrapolation in heterogeneous populations must explain how extrapolation can be possible even when such differences are present" (Steel 2007, 78-79).

[4] This does not mean that arguments from analogy are never affected by the extrapolator's circle (see Steel 2010 for a discussion). In the case of consciousness, however, the fact that standard systems are not necessarily assumed to be good models for non-standard systems seem to be enough to apply a charitable reading to the analysis.

This is a well-known problem in the social and life sciences: For example, the translational power of cancer research on animal studies to humans is limited (Mak, Evaniew, and Ghert 2014). Similarly, social policies and programs can fail when implemented in contexts that are partially different from the one in which the policy was previously (and successfully) implemented, as the case of the Tamil Nadu Integrated Nutrition Program[5] shows (Marchionni and Reijula 2019; Cartwright and Hardie 2012).

When it comes to consciousness, the problem is then: how can we be sure that the inevitable differences between neurotypical adult humans and non-standard systems are not differences that make a difference?

Typically, defenders of the analogical approach to the epistemological problem of the other minds (e.g., Hyslop and Jackson 1972; see Godfrey-Smith 2011 for a discussion) reply to this challenge by pointing out that the projectability of the property of interest is based on the fact that the property picks out a structural feature of reality, or, in other terms, a natural kind (i.e., a group of particulars bound together by how reality is, rather than by how observers think it is; Bird and Tobin 2008). If we drop a chicken's egg and observe that it breaks, we do not necessarily need to drop a seagull's egg, an ostrich's egg, and so on, to infer that those eggs will most probably break if dropped. The egg's fragility seems to be a property that depends on the egg's material constitution, and the egg's material constitution is a property reliably conserved across most, if not all, eggs. That is, natural kinds are supposed to be resistant

[5] This project was sponsored by the World Bank to reduce malnutrition in Indian communities by supplying food and by providing better nutritional knowledge to mothers. The success of the program was not replicated in Bangladesh, because of the differences in responsibility for the children's nutrition within the family. See Cartwright 2012 for a comprehensive discussion.

enough to differences between domains and contexts, so that properties of a member of the kind can be justifiably projected to other members of the kind.

This reply, based on the natural kind strategy, is also supposed to address another possible worry, namely that analogical reasoning for other consciousnesses is ultimately based on a sample of one population, and therefore cannot be informative. However, as Godfrey-Smith (2011) puts it, in the case of inductive inferences referring to natural kinds, "one instance of an F would be enough, in principle, if you picked the right case and analyzed it well. Ronald Reagan is supposed to have said 'once you've seen one redwood, you've seen them all'" (Godfrey-Smith 2011, 39).

The success of analogical reasoning to solve the problem of difference, and consequently the epistemological problem of other consciousness, thus seems to rest on whether the relationship between publicly observable properties and conscious properties is in fact a structural feature of reality or not; that is, if consciousness is indeed a natural kind.

Accordingly, to challenge the analogical inference, one could demonstrate that consciousness is not a natural kind. For example, it could be demonstrated that the concept 'consciousness' does not pick out any single phenomenon in reality, but rather a group of dissociable capacities and properties (Irvine 2012, 2017). However, most consciousness researchers implicitly operate under the assumption that consciousness is indeed a natural kind, as suggested by their attempts to uncover the neural basis of consciousness *as a unitary phenomenon* (e.g., Crick and Koch 1990; Melloni et al. 2021). Others embrace this view explicitly, and an active and ongoing research program has been leveraging this perspective (Shea 2012; Shea and Bayne 2010; Bayne and Shea 2020; Bayne et al. 2024; Mckilliam 2024).

Of course, this proclaimed consensus does not *guarantee* that consciousness is indeed a natural kind. To explain the core of the problem, let's go back to P1 of the argument schema introduced above: "D is a collection of data showing that there is a systematic and reliable correlation between publicly observable properties and consciousness in the SOURCE domain (i.e., neurotypical adult humans)".

The specific criteria needed to ensure that the correlation of interest tracks a natural kind will vary depending on which theory of natural kinds one endorses (for discussions, see Khalidi 2018; Boyd 2019). Yet, the minimal criterion for guaranteeing that the correlation can be validly projected is showing that it is not merely a spurious one: for example, this could be done by grounding the correlation on the presence of some mechanism that underlies the natural kind, and show that it generates similarity between members of the kind (but see Craver 2009 for a discussion). Of course, this would require identifying this mechanism, which might not be straightforward in the case of consciousness (Shea 2012; Bayne and Shea 2020). In any case, the link between publicly observable properties and consciousness should consistently and accurately reflect a structural feature of reality, not an observer-dependent artefact.

That is, the natural kind strategy should explain why the hypothesis of a direct, reliable, connection between consciousness and publicly observable properties is better than other explanations. In other words, the hypothesis that those publicly observable properties track a natural kind (i.e., consciousness) must be preferred to the hypothesis that the correlation between consciousness and those publicly observable properties is a spurious one. To do this, one could argue that the "natural kind hypothesis" is more parsimonious, or coheres better with background knowledge, than the "spurious correlation hypothesis". For example, following Sober (2000) (see also Millikan 1999), we could claim that consciousness is a

biological kind, and therefore is projectable to systems similar to us in terms of shared evolutionary history. In this case, the hypothesis that the relationship between consciousness and publicly observable properties is conserved in TARGET is *more parsimonious* than the hypothesis that analogous publicly observable properties are underwritten by conscious properties in one domain and unconscious properties in another domain. This is because the "consciousness hypothesis" requires only one character change (i.e., from creatures who do not have publicly observable properties correlated with consciousness to creatures who have such properties), while the "unconsciousness hypothesis" requires two character changes (i.e., from creatures who do not have those publicly observable properties to creatures who have those properties correlated with consciousness on the one hand and creatures who have those properties correlated with unconsciousness on the other hand). This strategy thus builds on parsimony considerations to explain why the hypothesis that a target system is conscious is better than alternative hypotheses.

The problem with analogical reasoning is that appealing to explanatory considerations of this sort, based on parsimony or coherence with background knowledge, pushes the limits of analogical reasoning by including elements that typically figure in abductive arguments. That is, analogical reasoning on its own cannot solve the problem; it must be combined with another type of reasoning. Specifically, it must be combined with IBE-based arguments, where the best explanation is justified precisely due to theoretical virtues like parsimony and coherence with background knowledge (Lipton 2001; McMullin 2008; Douglas 2013; Longino 1979; Psillos 2007).

To summarize: IBE-based arguments cannot solve the extrapolator's circle; analogical arguments can do that, but need to ensure that the projected property is a natural kind property in order to address the problem of difference. And to ensure that we are projecting a

genuine natural kind property, when we project consciousness from SOURCE to TARGET, analogical arguments must resort to explanatory considerations generally used by IBE-based arguments, which make them partially abductive: analogy does not seem to solve the problem of difference on its own.

Thus, both the IBE-based argument and the argument from analogy, individually taken, struggle with the challenges for successful extrapolations. For IBE-based reasoning to be successful, it needs to consider similarities between SOURCE and TARGET to solve the extrapolator's circle, while analogical reasoning needs to include explanatory considerations in order to solve the problem of difference. Both of them are incomplete, in this context.

To cope with this problem, we will now further systematize the approach already taken by some scholars in the field of consciousness science (e.g., Birch 2022; Barron and Klein 2016), suggesting that *analogical abduction* might solve this conundrum. We will provide a more systematic philosophical argument to justify this praxis, and claim that analogical reasoning and IBE-based reasoning are in fact complementary and could be merged to deliver a stronger form of reasoning to deal successfully with the epistemological problem of other consciousnesses.

4. Analogical Abduction

This is how Schurz (2008) defines analogical abduction:

> Here one abduces a partially new concept and at the same time new laws which connect this concept with given (empirical) concepts, in order to explain the given law-like phenomenon. The concept is only partly new because it is analogical to

> familiar concepts, and this is the way in which this concept was discovered. So analogical abduction is *driven* by analogy (Schurz 2008, 217).

The crucial point here is that abduction can confer justification to concepts posited and conjectured in the context of discovery, in virtue of their merits in explaining certain phenomena of interest. But the justification for positing such concepts is driven by analogical reasoning in the first place (Thagard 1988; Bartha 2019, 2010); namely, it is driven by the fact that the conjectured concepts are relevantly similar to some well-established concepts in our background knowledge.

We can formalize a general analogical abductive argument in the following way:

> P1. D is a collection of data about F-properties in SOURCE[6].
>
> P2. D* is a collection of data about F'-properties in TARGET.
>
> P3. There are relevant similarities between D and D* (From P1 & P2).
>
> Also,
>
> P4. We have good models that explanatorily link F-properties to G-properties in SOURCE.
>
> P5. The hypothesis H that G'-properties (which are similar to G-properties) occur in TARGET, which is formulated in virtue of P3, would explanatorily link F'-properties to G'-properties in TARGET.

[6] The argument presupposes that phenomena and their properties are causally related to the observed data, and that explanatory models (and associated hypotheses), although constructed upon data, are meant to explain those phenomena, not the data themselves (as argued by Bogen and Woodward 1988).

P6. H is better than any other hypothesis.

Therefore,

C. H is probably true (it is probably true that G'-properties occur in target).

If we apply this general schema to the epistemological problem of other consciousnesses, we have:

P1. D is a collection of data about publicly observable properties in SOURCE.

P2. D* is a collection of data about publicly observable properties in TARGET.

P3. There are relevant similarities between D and D* (From P1 & P2).

Also,

P4. We have good models that explanatorily link similar publicly observable properties to phenomenal properties in SOURCE.

P5. The hypothesis H that phenomenal properties are instantiated in TARGET (which we justifiably formulate because of P3) would explanatorily link publicly observable properties and phenomenal properties in TARGET, given that similar observable properties are explanatorily linked to phenomenal properties in SOURCE[7] (From P3 & P4).

P6. H is better than any other hypothesis.

Therefore,

C. It is probably true that phenomenal properties are instantiated in TARGET.

[7] To be precise, the hypothesis should posit that there are *phenomenal'* properties in TARGET, which are similar, but not identical to phenomenal properties in SOURCE. However, our primary focus here is on whether phenomenal properties are present or not, and because of this all we require is that the properties of interest be phenomenal.

The concept of 'consciousness-in-other-systems' ('TARGET-consciousness') is thus posited in virtue of the fact that we master, from the first-person perspective, the concept of 'consciousness-in-us' (SOURCE-consciousness), and that we possess a scientifically informed model of an explanatory relationship between consciousness and some publicly observable properties in us. Given such a reasonably well-established model, once we detect similar publicly observable properties in TARGET, it seems that the best hypothesis, in terms of parsimony and coherence with background knowledge[8], is that those properties are also related to consciousness in TARGET. This model of the relationship between consciousness and publicly observable properties in SOURCE does not need to be (or be derived from) a full-fledged theory of consciousness, but can also be derived from a more general framework that captures only some features of consciousness and some of the publicly observable properties related to it. What is relevant for the argument to remain abductive is that those features be *explanatorily* connected with publicly observable properties, and not purely correlated with them. In this sense, analogical abductive arguments can complement both "theory-heavy" (i.e., based on full-fledged and specific theories of consciousness – see Seth and Bayne 2022) and "theory-light" (Birch 2022) or "test-based" approaches (Bayne et al. 2024).

Thus, we have implemented Schurz's definition of analogical abduction on the epistemological problem of other consciousnesses: 'TARGET-consciousness' is the partially new concept, abduced in order to account for publicly observable properties in TARGET, but it is only *partially* new because it is analogical to the familiar concept of 'consciousness-in-us', or 'SOURCE-consciousness', and in virtue of this analogy the concept of TARGET-

[8] Ideally, the goodness of a scientific hypothesis should be systematized through a comprehensive taxonomy of explanatory virtues (see Keas 2018 for a discussion).

consciousness has been discovered. In the argument above, P1, P2 and P3 speak directly to analogical considerations, while P4, P5, and P6 speak to explanatory ones.

A final clarification pertains to the nature of the explanatory link mentioned in P4. This discussion interacts with the issue of what counts as relevant observable properties for consciousness. A first interpretation of the nature of this link is that F-properties are *causally explained* by G-properties. In the case of consciousness, F-properties could correspond to functional and/or behavioral properties, while G-properties would be conscious properties. Under this interpretation, the primary evidence for consciousness is given by functional and/or behavioral properties, and courtesy of analogical abductive arguments, consciousness in other systems would be posited to *causally explain* functional and behavioral properties observed in non-standard systems. For example, a theory like the global workspace theory (GNWT) (Mashour et al. 2020; Dehaene and Naccache 2001) posits that certain functions, like the ability to integrate and maintain information over time, can be performed only if information is broadcast into a global workspace that sustains and shares that information with many consumer systems. Accessibility to such global workspace just *is* consciousness. Those functional abilities can be evidenced by mental chaining of operations (Sackur and Dehaene 2009) or global violation detection (Bekinschtein et al. 2009), just to cite a few, because these functions are causally connected to consciousness and cannot be performed unconsciously. Thus, if a target system is able to perform tasks that show it possesses such abilities (e.g., it detects a global variation in a sequence of auditory stimuli), then GNWT proponents would be entitled to formulate analogical abductive arguments based on the fact that those behaviors are *causally* explained by accessibility of information to a global workspace, i.e., by consciousness. Similarly, Birch's "theory-light" approach assumes that the identification of some cognitive functions that are facilitated by consciousness suffices to

extrapolate from the human case to non-human animals. Even without endorsing a specific theory, this approach can still take advantage of analogical abductive arguments by positing that the nature of the explanatory link between consciousness and a cluster of publicly observable properties (e.g., unlimited associative learning; Birch, Ginsburg and Jablonka 2020) must be causal: consciousness is assumed to be the cause of certain functions and behaviors in us; if similar functions and behaviors are observed in certain target systems, one could formulate an analogical abductive argument to justify the attribution of consciousness to those systems. Thus, independently of whether one operates under the tenets of a specific theory or not, if functional/behavioral properties are taken to be the primary evidence for consciousness, then the nature of the explanatory link in the analogical abductive argument will likely be causal.

But this causal explanatory link does not hold if *implementation or mechanistic properties* (e.g., neurobiological properties) are taken to be primary evidence for consciousness, because implementation properties are not causally explained by conscious properties. In this case, we can take advantage of a type of abduction that Harman (1986, 68) and Lipton (2004) have called "inference *from* the best explanation". In this instance, we do not observe what *could be explained* by the hypothesis, but rather we observe what *could explain* a fact, or a property, that we are licensed to posit in virtue of that observation. With Lipton's example, from the observation that it is cold outside, I am justified to infer that the car will not start: the observation (i.e., "it is cold outside") is not explained by the hypothesis that the car will not start. Rather, as Lipton puts it, given relevant background knowledge, I am justified to infer that the car will not start from the observation that it is cold outside because if it were true that the car will not start, the cold would be the best explanation for it (Lipton 2004, 64).

In the inference *from* the best explanation, the observation is the explanation[9], not the conjecture. Thus, in the case of other consciousnesses, inference from the best explanation could license the conclusion that publicly observed properties explain the consciousness of the target system because if it were true that the target system is conscious, those properties would be the best explanation of that fact. Therefore, if we take the primary evidence for consciousness to be neurobiological/mechanistic properties, we can posit conscious properties (G'-properties) in TARGET because, given what we know about the neural underpinnings of consciousness in us (i.e., in SOURCE), if it were true that the system of interest in TARGET is conscious, the observed neurobiological/mechanistic properties in TARGET (F'-properties) would be the best explanation of that fact.

For example, the Recurrent Processing Theory (RPT) (Lamme 2006; Lamme and Roelfsema 2000) maintains that consciousness corresponds to the implementation of information processing on local feedback loops (involving synaptic plasticity) in sensory areas of the brain. If a target system displayed information processing implemented through feedback loops, RPT proponents could formulate an analogical abductive argument to justify the attribution of consciousness to the target system. The explanatory link in their P4 would be constitutive or mechanistic, insofar as the explanatory model they rely on (i.e., RPT) is based on a *constitutive* relationship between consciousness and the explanatorily relevant publicly observable property (i.e., feedback loops). Thus, if mechanistic/implementational properties are taken to be the primary evidence for consciousness, then the nature of the explanatory

[9] For Lipton, the nature of this explanation should be causal (Lipton 2001). We are not making that assumption here.

link between consciousness and publicly observable properties will likely be constitutive or mechanistic (Craver 2007) rather than causal (for discussion, see Prasetya 2021).

Thus, different types of explanatory links between phenomenal properties and publicly observable properties can be given, depending on what researchers consider to be the relevant evidence for consciousness. Analogical abductive arguments will accordingly take different specific forms[10].

The analogical abductive approach is thus highly promising, and can be easily applied to existing theories and accounts of consciousness. Yet, it is not devoid of limitations, which we address in the next section, where we critically examine this strategy and its potential.

5. The Prospects and Limitations of Analogical Abduction

Despite its virtues, analogical abduction is not a silver bullet, and faces several problems. As a starting point, we focus on the above-mentioned challenges for successful extrapolations, which analogical abduction manages to meet, and then continue to more contentious issues.

---

[10] The conclusion of these arguments would be justified only if we had good knowledge of the functional and behavioral profile associated with consciousness on the one hand (for arguments based on explanatory links of *causal* nature), or its mechanistic underpinnings on the other hand (for arguments based on explanatory links of *constitutive/mechanistic* nature). However, the current status of consciousness science does not provide us with good knowledge of either the functional profile of consciousness or its neural mechanisms (see, e.g., Francken et al. 2022; Yaron et al. 2022). This limitation reduces the inferential strength of analogical abductive arguments. We will elaborate on this further in Section 4.

First, the "extrapolator's circle", namely the problem of explaining why we are justified in believing that a TARGET system is conscious without already assuming that the best explanation for SOURCE-consciousness works in TARGET. As opposed to the IBE-based argument, analogical abduction avoids this problem because the link between observable properties and consciousness in TARGET is not simply assumed; rather, it is justified *in virtue of the similarity between properties in TARGET and certain properties in SOURCE* (thus, this strategy goes beyond the IBE-based one by relying on analogical reasoning). Again, analogical considerations drive the abduction process, and that is why the TARGET system is not considered conscious simply in virtue of the assumption that our best explanation for SOURCE-consciousness works for TARGET too. Rather, the hypothesis that the target system might be conscious is grounded on the similarity between TARGET and SOURCE. The key point is that this similarity justifies the *formulation* of the hypothesis that the target system might be conscious – a hypothesis that could not be justifiably formulated without these analogical considerations.

Second, the problem of difference: how can we determine whether a system is similar enough to us to justify extrapolations about its consciousness? A possible solution can be based on Gentner's structure-mapping theory (Gentner 1983), which influenced Guala's thesis that extrapolation is possible only when the source and target systems "belong to structurally similar mechanisms" (Guala 2005, 180), as well as Steel's notion of comparative process tracing (Steel 2007; Guala 2010; Steel 2010). According to this approach, similarity is structurally grounded: extrapolations are justified insofar as the mechanistic processes or the properties in TARGET have the same structure of the relevant processes or properties in SOURCE. Translated to consciousness science, this would mean that extrapolations about consciousness are justified only when there is a *structure-preserving mapping* between the

consciousness-related properties in SOURCE and the properties observed in TARGET. This means that if we had a compelling model of SOURCE-consciousness that explanatorily relates consciousness to a (causal or constitutive) structure of publicly observable properties (thereby going beyond the analogical strategy and relying on aspects of the IBE-based one), we could project consciousness to any system that exhibits the same structure of publicly observable properties, independently of all the other possible differences.

Analogical abductions thus inherit the inferential relevance of explanatory considerations from the IBE-based argument, and the importance of structural similarity for the projectability of the property of interest from the argument from analogy.

Admittedly however, the analogical abductive strategy is limited, since structural similarity is a property that comes in degrees, and it is unclear what level of similarity suffices to justify extrapolations. In order to make the extrapolative leap, we should define a "similarity threshold" above which the inference is justified. Some have suggested that this threshold might be based on a *cluster* of similar properties between SOURCE and TARGET (Birch 2022) but this still requires a definition of the minimal size of the cluster that justifies the extrapolative leap (Shevlin 2021). Moreover, although we consider this strategy to be promising for extrapolating consciousness to biological creatures, it might be more difficult to apply to artificial systems. This is mainly because of three reasons: in the case of artificial consciousness, (I) we cannot rely on evolutionary similarities; (II) metaphysical debates concerning the substrate neutrality of consciousness are more prominent (Shiller 2024; Seth 2024); and (III) our antecedent knowledge that the cluster of markers associated with consciousness in humans has been deliberately *designed* to be displayed by an artificial entity might undermine the view that this cluster tracks consciousness in this domain.

A third problem for the analogical abduction strategy is that it presently has a limited 'epistemic force'. In the philosophical literature (Calzavarini and Cevolani 2022; Schurz 2008), abductive arguments are considered as *strong* if they justify the acceptance of a hypothesis as true, while they are considered as *weak* if they just select hypotheses as interesting conjectures that require further empirical testing.

Given the current disagreement in consciousness science and the lack of a well-established, and specific, theory of consciousness (Francken et al. 2022; Yaron et al. 2022), analogical abductive arguments for non-standard systems seem to be *weak* (Baetu 2024). If so, they can only be used to justify *formulating the hypothesis* that a certain target system is conscious, rather than *accepting the hypothesis* that it actually is conscious. Since we already relied on the relevant evidence to build the analogical abductive argument and formulate the hypothesis that the system is conscious, we seem to lack the methodological basis for passing from hypothesis-formulation to hypothesis-acceptance. Thus, until a better understanding of consciousness is available, this approach can only offer a partial solution for the epistemological problem of other consciousnesses.

A possible way to establish such a methodological basis might lie in the iterative natural kind approach (Bayne et al. 2024; Mckilliam 2024), since its iterative nature can allow a gradual increase in our confidence in whether the target system is conscious (see also Baetu 2024). However, it is not clear whether this strategy can deliver *strong* analogical abductive inferences, rather than just *less weak* inferences.

In the meantime, analogical abductive arguments could still be helpful in delivering weakly justified extrapolations. In several decision-making contexts, especially those related to substantial ethical and societal implications, it is reasonable to lower the evidential bar and

enact evidence-based policies that can preempt harm, even if the evidence is only partial. As Steel puts it, “policy [should] not be susceptible to paralysis by scientific uncertainty” (Steel 2013, 321 – cited in Birch 2017. See also Birch 2023 and Johnson 2016). Analogical abductive arguments can serve as the tools to navigate the uncertainty about consciousness in non-standard systems and to provide attributions of consciousness that, albeit weak, can still be sufficient for informed decision-making.

To summarize, analogical abductive arguments are currently facing a speed/accuracy trade-off: they can deliver strongly justified and accurate conclusions either via an established theory of consciousness or by using the iterative natural kind strategy; both options, however, require a long process that will not likely be completed in the near future. On the other hand, consciousness science is already needed to inform decision-making and regulations about various non-standard systems. In the short run, analogical abductive arguments can provide some degree of justification for attributions of consciousness to these systems, but since these attributions can be only weakly justified, the risk of inaccurate attributions is high.

6. Conclusion

In this paper, we maintain that arguments based on analogical abduction capture and systematize the reasoning strategy employed by many consciousness researchers interested in attributing consciousness to non-standard systems and that this strategy is indeed the most promising for extrapolating the presence of consciousness in such systems. Analogical abductive arguments do better than standard analogical arguments and IBE-based arguments with respect to the two challenges that any extrapolative inference must satisfy, namely the extrapolator’s circle and the problem of difference. However, further research is needed to allow analogical abductive arguments to overcome some of the limitations they are still

facing. For example, it is not clear what degree of similarity along consciousness-relevant dimensions is sufficient to justify strong abductions. Moreover, it seems possible that different degrees of similarity will be required, depending on the type of conclusion we are interested in: inferring that a system is conscious (rather than not) might require a lower degree of similarity than a conclusion about what the system is conscious of.

Other problems, although not directly related to the structure of analogical abductive arguments, might limit their applicability: the current status of consciousness science still does not provide a clear view on what type of publicly observable properties are relevant to consciousness specifically, and consensus on the best theory of SOURCE-consciousness is far from being near.

Thus, the applicability of analogical abductions is currently limited. However, analogical abductions still provide the best available justificatory option for inferring consciousness in non-standard systems, and can be useful to derive weakly justified conclusions that can inform practical decision-making. We have argued that analogical abductive arguments provide the blueprint for justified attributions of consciousness to non-standard systems. This is important because a clarification of the argumentative structure underlying our ascriptions of consciousness can help consciousness scholars assess the soundness of existing arguments for attributing consciousness in non-standard systems, as well as formulate stronger arguments of this type.

References

Albantakis, Larissa, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjørn E. Juel, Shuntaro Sasai, Keiko Fujii, Isaac David, Jeremiah Hendren, Jonathan P. Lang, and Giulio Tononi. 2023. "Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms." *PLOS Computational Biology* 19 (10): e1011465. https://doi.org/10.1371/journal.pcbi.1011465.

Andrews, Kristin. 2024. "“All animals are conscious”: Shifting the null hypothesis in consciousness science." *Mind & Language*. https://doi.org/https://doi.org/10.1111/mila.12498.

Avramides, Anita. 2000. *Other Minds*. London: Routledge.

Baetu, Tudor M. 2024. "Extrapolating animal consciousness." *Studies in History and Philosophy of Science* 104: 150-159. https://doi.org/10.1016/j.shpsa.2024.03.001.

Barron, Andrew B., and Colin Klein. 2016. "What insects can tell us about the origins of consciousness." *Proceedings of the National Academy of Sciences* 113 (18): 4900-4908. https://doi.org/doi:10.1073/pnas.1520084113.

Bartha, Paul. 2010. *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press.

---. 2019. Analogy and Analogical Reasoning. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Bayne, Tim. 2018. "On the axiomatic foundations of the integrated information theory of consciousness." *Neurosci Conscious* 2018 (1): niy007. https://doi.org/10.1093/nc/niy007.

Bayne, Tim, Joel Frohlich, Rhodri Cusack, Julia Moser, and Lorina Naci. 2023. "Consciousness in the cradle: on the emergence of infant experience." *Trends in Cognitive Sciences* 27 (12): 1135-1149. https://doi.org/10.1016/j.tics.2023.08.018.

Bayne, Tim, Anil K. Seth, and M. Massimini. 2020. "Are There Islands of Awareness?" *Trends in Neurosciences* 43 (1): 6-16. https://doi.org/10.1016/j.tins.2019.11.003.

Bayne, Tim, Anil K. Seth, Marcello Massimini, Joshua Shepherd, Axel Cleeremans, Stephen M. Fleming, Rafael Malach, Jason B. Mattingley, David K. Menon, Adrian M. Owen,

Megan A. K. Peters, Adeel Razi, and Liad Mudrik. 2024. "Tests for consciousness in humans and beyond." *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2024.01.010.

Bayne, Tim, and Nicholas Shea. 2020. "Consciousness, Concepts and Natural Kinds." *Philosophical Topics* 48 (1): 65-83. https://doi.org/https://www.jstor.org/stable/48628586.

Bekinschtein, T.A., S. Dehaene, B. Rohaut, F. Tadel, L. Cohen, and L. Naccache. 2009. "Neural signature of the conscious processing of auditory regularities." *Proceedings of the National Academy of Sciences* 106 (5): 1672-1677. https://doi.org/10.1073/pnas.0809667106.

Birch, Jonathan. 2017. "Animal Sentience and the Precautionary Principle." *Animal Sentience* 2: 16(1). https://doi.org/10.51291/2377-7478.1200.

---. 2022. "The search for invertebrate consciousness." *Noûs* 56 (1): 133-153. https://doi.org/https://doi.org/10.1111/nous.12351.

---. 2023. "Medical AI, inductive risk and the communication of uncertainty: the case of disorders of consciousness." *Journal of Medical Ethics*: jme-2023-109424. https://doi.org/10.1136/jme-2023-109424.

---. 2024. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.

Birch, Jonathan, and H. Browning. 2021. "Neural Organoids and the Precautionary Principle." *Am J Bioeth* 21 (1): 56-58. https://doi.org/10.1080/15265161.2020.1845858.

Birch, Jonathan, Simona Ginsburg, and Eva Jablonka. 2020. "Unlimited Associative Learning and the origins of consciousness: a primer and some predictions." *Biology & Philosophy* 35 (6): 56. https://doi.org/10.1007/s10539-020-09772-0.

Bird, Alexander, and Emma Tobin. 2008. "Natural kinds." In *Standford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Block, Ned. 2002. "The Harder Problem of Consciousness." *The Journal of Philosophy* 99 (8): 391-425. https://doi.org/10.2307/3655621.

---. 2007. "Consciousness, accessibility, and the mesh between psychology and neuroscience." *Behavioral and Brain Sciences* 30 (5-6): 481-499. https://doi.org/10.1017/S0140525X07002786.

Bogen, James, and James Woodward. 1988. "Saving the phenomena." *Philosophical Review* 97 (3): 303-352. https://doi.org/https://doi.org/10.2307/2185445.

Boyd, Richard. 2019. "Rethinking natural kinds, reference and truth: towards more correspondence with reality, not less." *Synthese* 198 (Suppl 12): 2863-2903. https://doi.org/https://doi.org/10.1007/s11229-019-02138-4.

Butlin, P., R. Long, E. Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, G. Deane, Stephen Fleming, Chris D. Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, M. Michel, L. Mudrik, Megan A. K. Peters, Eric Schwitzgebel, J. Simon, and R. VanRullen. 2023. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." *arXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2308.08708.

Calzavarini, Fabrizio, and Gustavo Cevolani. 2022. "Abductive reasoning in cognitive neuroscience: weak and strong reverse inference." *Synthese* 200 (2): 70. https://doi.org/10.1007/s11229-022-03585-2. https://doi.org/10.1007/s11229-022-03585-2.

Carruthers, Peter. 2019. *Human and Animal Minds: The Consciousness Questions Laid to Rest*. Oxford University Press.

Cartwright, Nancy. 1994. *Nature's Capacities and Their Measurement*. Oxford University Press.

---. 2012. "Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps." *Philosophy of Science* 79 (5): 973-989. https://doi.org/10.1086/668041.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing It Better*. Oxford University Press.

Chalmers, David J. 1996. *The conscious mind : in search of a fundamental theory.Philosophy of mind series*. New York: Oxford University Press.

---. 2023. "Could a large language model be conscious?".

Ciaunica, Anna, Adam Safron, and Jonathan Delafield-Butt. 2021. "Back to square one: the bodily roots of conscious experiences in early life." *Neuroscience of Consciousness* 2021 (2). https://doi.org/10.1093/nc/niab037.

Craver, Carl F. 2002. "Structures of Scientific Theories." In *The Blackwell Guide to the Philosophy of Science*, edited by Peter Machamer and Michael Silberstein, 55–79. Blackwell.

---. 2007. *Explaining the Brain*. New York: Oxford University Press.

---. 2009. "Mechanisms and natural kinds." *Philosophical Psychology* 22 (5): 575-594. https://doi.org/10.1080/09515080903238930.

Crick, F., and Christof Koch. 1990. "Toward a neurobiological theory of consciousness." *Seminars in the Neurosciences* 2: 263-275. https://doi.org/https://doi.org/101584582X469.

Croxford, James, and Tim Bayne. 2024. "The Case Against Organoid Consciousness." *Neuroethics* 17 (1): 13. https://doi.org/10.1007/s12152-024-09548-3. https://doi.org/10.1007/s12152-024-09548-3.

Dehaene-Lambertz, Ghislaine. 2024. "Perceptual Awareness in Human Infants: What is the Evidence?" *Journal of Cognitive Neuroscience*: 1-11. https://doi.org/10.1162/jocn_a_02149.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider. 2017. "What is consciousness, and could machines have it?" *Science* 358 (6362): 486-492. https://doi.org/doi:10.1126/science.aan8871.

Dehaene, Stanislas, and Lionel Naccache. 2001. "Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework." *Cognition* 79 (1): 1-37. https://doi.org/https://doi.org/10.1016/S0010-0277(00)00123-2.

Douglas, Heather. 2013. "The Value of Cognitive Values." *Philosophy of Science* 80 (5): 796-806. https://doi.org/https://doi.org/10.1086/673716.

Dung, Leonard. 2022. "Assessing tests of animal consciousness." *Consciousness and Cognition* 105: 103410. https://doi.org/https://doi.org/10.1016/j.concog.2022.103410.

---. 2023. "Tests of Animal Consciousness are Tests of Machine Consciousness." *Erkenntnis*. https://doi.org/10.1007/s10670-023-00753-9.

Elamrani, A., and R. V. Yampolskiy. 2019. "Reviewing Tests for Machine Consciousness." *Journal of Consciousness Studies* 26 (5-6): 35-64. https://www.ingentaconnect.com/content/imp/jcs/2019/00000026/f0020005/art00002.

Francken, Jolien C, Lola Beerendonk, Dylan Molenaar, Johannes J Fahrenfort, Julian D Kiverstein, Anil K. Seth, and Simon van Gaal. 2022. "An academic survey on theoretical foundations, common assumptions and the current state of consciousness science." *Neuroscience of Consciousness* 2022 (1). https://doi.org/10.1093/nc/niac011.

Frohlich, Joel, Tim Bayne, Julia S. Crone, Alessandra DallaVecchia, Asger Kirkeby-Hinrup, Pedro A. M. Mediano, Julia Moser, Karolina Talar, Alireza Gharabaghi, and Hubert Preissl. 2023. "Not with a “zap” but with a “beep”: Measuring the origins of perinatal experience." *NeuroImage* 273: 120057. https://doi.org/https://doi.org/10.1016/j.neuroimage.2023.120057.

Gentner, Dedre. 1983. "Structure-mapping: A theoretical framework for analogy." *Cognitive Science* 7 (2): 155-170. https://doi.org/https://doi.org/10.1016/S0364-0213(83)80009-3.

Godfrey-Smith, Peter. 2011. "Induction, Samples, and Kinds." In *Carving Nature at its Joints: Natural Kinds in Metaphysics and Science*, edited by Michael O'Rourke, Joseph Keim Campbell and Matthew H. Slater, 33-52. MIT Press.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge University Press.

---. 2010. "Extrapolation, Analogy, and Comparative Process Tracing." *Philosophy of Science* 77 (5): 1070-1082. https://doi.org/10.1086/656541.

Halina, Marta, David Harrison, and Colin Klein. 2022. "Evolutionary Transition Markers and the Origins of Consciousness." *Journal of Consciousness Studies* 29 (3-4): 62-77. https://doi.org/10.53765/20512201.29.3.077.

Hameroff, Stuart, and Alysson R. Muotri. 2020. "Testing for consciousness in cerebral organoids." *Trends in Cell & Molecular Biology* 15: 43-57.

Harman, Gilbert. 1986. *Change in View: Principles of Reasoning*. Vol. 1: MIT Press.

Hesse, Mary. 1965. "Models and Analogies in Science." *British Journal for the Philosophy of Science* 16 (62): 161-163.

Heyes, Cecilia. 2008. "Beast machines? Questions of animal consciousness." In *Frontiers of consciousness*, edited by Lawrence Weiskrantz and Martin Davies, 259--274. Oxford University Press.

Hiddleston, Eric. 2005. "Causal Powers." *The British Journal for the Philosophy of Science* 56 (1): 27-59. https://doi.org/10.1093/phisci/axi102.

Hildt, Elisabeth. 2022. "The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns." *AJOB Neuroscience*: 1-14. https://doi.org/10.1080/21507740.2022.2148773.

Hyslop, A. 1995. "The Analogical Inference to Other Minds." In *Other Minds*, edited by Alec Hyslop, 41-70. Dordrecht: Springer Netherlands.

Hyslop, A., and F. C. Jackson. 1972. "The Analogical Inference to Other Minds." *American Philosophical Quarterly* 9 (2): 168-176. http://www.jstor.org/stable/20009435.

Irvine, Elizabeth. 2012. *Consciousness as a scientific concept: a philosophy of science perspective*. Vol. 5 *Studies in Brain and Mind*: Springer.

---. 2017. "Explaining What?" *Topoi* 36 (1): 95-106. https://doi.org/http://dx.doi.org/10.1007/s11245-014-9273-4.

Jeziorski, Jacob, Reuven Brandt, John H. Evans, Wendy Campana, Michael Kalichman, Evan Thompson, Lawrence Goldstein, Christof Koch, and Alysson R. Muotri. 2023. "Brain organoids, consciousness, ethics and moral status." *Seminars in Cell & Developmental Biology* 144: 97-102. https://doi.org/https://doi.org/10.1016/j.semcdb.2022.03.020.

Johnson, L. Syd M. 2016. "Inference and Inductive Risk in Disorders of Consciousness." *American Journal of Bioethics Neuroscience* 7 (1): 35-43. https://doi.org/https://doi.org/10.1080/21507740.2016.1150908.

Kanai, Ryota, and Ippei Fujisawa. 2024. "Toward a universal theory of consciousness." *Neuroscience of Consciousness* 2024 (1). https://doi.org/10.1093/nc/niae022.

Kazazian, Karnig, Brian L. Edlow, and Adrian M. Owen. 2024. "Detecting awareness after acute brain injury." *The Lancet Neurology* 23 (8): 836-844. https://doi.org/10.1016/S1474-4422(24)00209-6.

Keas, Michael N. 2018. "Systematizing the theoretical virtues." *Synthese* 195 (6): 2761-2793. https://doi.org/10.1007/s11229-017-1355-6.

Khalidi, Muhammad Ali. 2018. "Natural kinds as nodes in causal networks." *Synthese* 195 (4): 1379-1396. https://doi.org/https://doi.org/10.1007/s11229-015-0841-y.

LaFollette, H., and N. Shanks. 1996. *Brute Science: Dilemmas of Animal Experimentation*. Routledge.

Lamme, V.A.F. 2006. "Towards a true neural stance on consciousness." *Trends Cogn Sci* 10 (11): 494-501. https://doi.org/10.1016/j.tics.2006.09.001.

Lamme, V.A.F., and P.R. Roelfsema. 2000. "The distinct modes of vision offered by feedforward and recurrent processing." *Trends in Neuroscience* 23 (11): 571-579. https://doi.org/10.1016/S0166-2236(00)01657-X.

Lau, H., and M. Michel. 2019. "On the dangers of conflating strong and weak versions of a theory of consciousness.". https://doi.org/https://doi.org/10.31234/osf.io/hjp3s.

Lavazza, A., and M. Massimini. 2018. "Cerebral organoids: ethical issues and consciousness assessment." *J Med Ethics* 44 (9): 606-610. https://doi.org/10.1136/medethics-2017-104555.

Levy, Neil. 2024. "Consciousness Ain't All That." *Neuroethics* 17 (2): 1-14. https://doi.org/https://doi.org/10.1007/s12152-024-09559-0.

Lipton, Peter. 2001. "What Good is an Explanation?" In *Explanation: Theoretical Approaches and Applications*, edited by Giora Hon and Sam S. Rakover, 43-59. Dordrecht: Springer Netherlands.

---. 2004. *Inference to the Best Explanation*. Routledge.

Longino, Helen E. 1979. "Evidence and hypothesis: An analysis of evidential relations." *Philosophy of Science* 46 (1): 35-56. https://doi.org/10.1086/288849.

Mak, I. W., N. Evaniew, and M. Ghert. 2014. "Lost in translation: animal models and clinical trials in cancer treatment." *Am J Transl Res* 6 (2): 114-8.

Marchionni, C., and S. Reijula. 2019. "What is mechanistic evidence, and why do we need it for evidence-based policy?" *Stud Hist Philos Sci* 73: 54-63. https://doi.org/10.1016/j.shpsa.2018.08.003.

Mashour, G. A., P. Roelfsema, Jean-Pierre Changeux, and Stanislas Dehaene. 2020. "Conscious Processing and the Global Neuronal Workspace Hypothesis." *Neuron* 105 (5): 776-798. https://doi.org/10.1016/j.neuron.2020.01.026.

Mckilliam, Andy. 2024. "Natural kind reasoning in consciousness science: An alternative to theory testing." *Noûs* n/a (n/a). https://doi.org/https://doi.org/10.1111/nous.12526.

McMullin, Ernan. 2008. "The virtues of a good theory." In *The Routledge Companion to Philosophy of Science*, edited by Martin Curd and Stathis Psillos. Routledge.

Mediano, P. A. M., Fernando E. Rosas, Daniel Bor, Anil K. Seth, and A. B. Barrett. 2022. "The strength of weak integrated information theory." *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.04.008.

Melloni, Lucia, Liad Mudrik, Michael Pitts, and Christof Koch. 2021. "Making the hard problem of consciousness easier." *Science* 372 (6545): 911-912. https://doi.org/doi:10.1126/science.abj3259.

Melnyk, Andrew. 1994. "Inference to the best explanation and other minds." *Australasian Journal of Philosophy* 72 (4): 482-491. https://doi.org/10.1080/00048409412346281.

Merker, Bjorn, Kenneth Williford, and David Rudrauf. 2021. "The Integrated Information Theory of consciousness: A case of mistaken identity." *Behavioral and Brain Sciences*: 1-72. https://doi.org/10.1017/S0140525X21000881.

Millikan, Ruth Garrett. 1999. "Historical Kinds and the "Special Sciences". *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 95 (1/2): 45-65. http://www.jstor.org/stable/4320948.

Moser, J., F. Schleger, M. Weiss, K. Sippel, L. Semeia, and H. Preissl. 2021. "Magnetoencephalographic signatures of conscious processing before birth." *Dev Cogn Neurosci* 49: 100964. https://doi.org/10.1016/j.dcn.2021.100964.

Mudrik, Liad, Melanie Boly, Stanislas Dehaene, Stephen M. Fleming, Victor Lamme, Anil Seth, and Lucia Melloni. 2025. "Unpacking the complexities of consciousness: Theories and reflections." *Neuroscience & Biobehavioral Reviews* 170: 106053. https://doi.org/https://doi.org/10.1016/j.neubiorev.2025.106053.

Mudrik, Liad, Myrto Mylopoulos, Niccolo Negro, and Aaron Schurger. 2023. "Theories of consciousness and a life worth living." *Current Opinion in Behavioral Sciences* 53: 101299. https://doi.org/https://doi.org/10.1016/j.cobeha.2023.101299.

Negro, Niccolo, and Liad Mudrik. 2025. “Testing for consciousness beyond consensus cases”. *PhilSci Archive*

Owen, Adrian M., Martin R. Coleman, Melanie Boly, Matthew H. Davis, Steven Laureys, and John D. Pickard. 2006. "Detecting Awareness in the Vegetative State." *Science* 313 (5792): 1402-1402. https://doi.org/10.1126/science.1130197.

Owen, Matthew, Zirui Huang, Catherine Duclos, Andrea Lavazza, Matteo Grasso, and Anthony G. Hudetz. 2023. "Theoretical Neurobiology of Consciousness Applied to Human Cerebral Organoids." *Cambridge Quarterly of Healthcare Ethics*: 1-21. https://doi.org/10.1017/S0963180123000543.

Pargetter, Robert. 1984. "The scientific inference to other minds." *Australasian Journal of Philosophy* 62 (2): 158-163. https://doi.org/10.1080/00048408412341341.

Passos-Ferreira, Claudia. 2023. "Are Infants Conscious?" *Philosophical Perspectives* 37 (1): 308-329. https://doi.org/https://doi.org/10.1111/phpe.12192.

---. 2024. "Can we detect consciousness in newborn infants?" *Neuron* 112 (10): 1520-1523. https://doi.org/10.1016/j.neuron.2024.04.024.

Prasetya, Yunus. 2021. "Which Models of Scientific Explanation are (In)Compatible with IBE?" *The British Journal for the Philosophy of Science* 0 (ja): null. https://doi.org/10.1086/715203.

Psillos, Stathis. 2002. "Simply the best: A case for abduction." In *Computational Logic: Logic Programming and Beyond : Essays in Honour of Robert A. Kowalski, Part Ii*, 83-93. Springer Berlin.

---. 2007. "The Fine Structure of Inference to the Best Explanation." *Philosophy and Phenomenological Research* 74 (2): 441-448. https://doi.org/https://doi.org/10.1111/j.1933-1592.2007.00030.x.

Sackur, Jérôme, and Stanislas Dehaene. 2009. "The cognitive architecture for chaining of two mental operations." *Cognition* 111 (2): 187-211. https://doi.org/https://doi.org/10.1016/j.cognition.2009.01.010.

Schneider, Susan. 2019. *Artificial You: AI and the Future of Your Mind*. Princeton University Press.

Schurz, G. 2008. "Patterns of abduction." *Synthese* 164 (2): 201-234. https://doi.org/10.1007/s11229-007-9223-4.

Seth, Anil K. 2024. "Conscious artificial intelligence and biological naturalism." *PsyArXiv*. https://doi.org/https://osf.io/preprints/psyarxiv/tz6an_v1.

Seth, Anil K., and Tim Bayne. 2022. "Theories of consciousness." *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-022-00587-4.

Shea, Nicholas. 2012. "Methodological Encounters with the Phenomenal Kind." *Philosophy and Phenomenological Research* 84 (2): 307-344. https://doi.org/https://doi.org/10.1111/j.1933-1592.2010.00483.x.

Shea, Nicholas, and Tim Bayne. 2010. "The Vegetative State and the Science of Consciousness*." *The British Journal for the Philosophy of Science* 61 (3): 459-484. https://doi.org/10.1093/bjps/axp046.

Shepherd, J. 2018. *Consciousness and Moral Status*. Taylor & Francis.

Shevlin, Henry. 2021. "Non-human consciousness and the specificity problem: A modest theoretical proposal." *Mind & Language* 36 (2): 297-314. https://doi.org/https://doi.org/10.1111/mila.12338.

Shiller, Derek. 2024. "Functionalism, integrity, and digital consciousness." *Synthese* 203 (2): 47. https://doi.org/10.1007/s11229-023-04473-z.

Siewert, Charles P. 1998. *The Significance of Consciousness*. Princeton University Press.

Sills, Jennifer, Olivia Carter, Jakob Hohwy, Jeroen van Boxtel, Victor Lamme, Ned Block, Christof Koch, and Naotsugu Tsuchiya. 2018. "Conscious machines: Defining questions." *Science* 359 (6374): 400-400. https://doi.org/doi:10.1126/science.aar4163.

Sober, Elliott. 2000. "Evolution and the Problem of Other Minds." *Journal of Philosophy* 97 (7): 365. https://doi.org/10.2307/2678410.

Steel, Daniel. 2007. *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford University Press.

---. 2010. "A New Approach to Argument by Analogy: Extrapolation and Chain Graphs." *Philosophy of Science* 77 (5): 1058-1069. https://doi.org/10.1086/656543.

---. 2013. "The Precautionary Principle and the Dilemma Objection." *Ethics, Policy & Environment* 16 (3): 321-340. https://doi.org/10.1080/21550085.2013.844570.

Thagard, Paul. 1988. *Computational philosophy of science*.*Computational philosophy of science.* Cambridge, MA, US: The MIT Press.

Tononi, Giulio, M. Boly, M. Massimini, and Christof Koch. 2016. "Integrated information theory: from consciousness to its physical substrate." *Nat Rev Neurosci* 17 (7): 450-61. https://doi.org/10.1038/nrn.2016.44.

Tye, Michael. 2017. *Tense bees and shell-shocked crabs: Are animals conscious?*. New York, NY, US: Oxford University Press. doi:10.1093/acprof:oso/9780190278014.001.0001.

Usher, M., Niccolò Negro, Hilla Jacobson, and N. Tsuchiya. 2023. "When philosophical nuance matters: safeguarding consciousness research from restrictive assumptions." *Frontiers in Psychology* 14. https://doi.org/10.3389/fpsyg.2023.1306023.

Veit, Walter. 2022. "Towards a Comparative Study of Animal Consciousness." *Biological Theory* 17 (4): 292-303. https://doi.org/10.1007/s13752-022-00409-x.

Yaron, I., L. Melloni, M. Pitts, and L. Mudrik. 2022. "The ConTraSt database for analysing and comparing empirical studies of consciousness theories." *Nat Hum Behav*. https://doi.org/10.1038/s41562-021-01284-5.

Acknowledgments

We wish to thank all members of the Mudrik Lab at Tel Aviv University for their comments on previous drafts of this manuscript, in particular Gennadiy Belonosov, Rony Hirschhorn, Maor Schreiber, and Amir Tal. NN thanks Andy McKilliam for reading and providing comments on an early version of this work. A previous version of this paper was presented at ASSC27, where we received very helpful comments and feedback from the audience. NN wishes to acknowledge the support of the Azrieli Foundation.