(Dis)confirming Theories of Consciousness and Their Predictions: Towards a Lakatosian Consciousness Science

Niccolò Negro

School of Psychological Sciences, Tel Aviv University, Tel Aviv-Yafo, Israel

niccolonegro@tauex.tau.ac.il; niccolo.negro.research@gmail.com

Highlights

- Consciousness science is witnessing a substantial empirical acceleration thanks to adversarial collaborations.
- Consciousness science needs to interact with confirmation theory to better understand the (dis)confirmatory value of empirical evidence for theories of consciousness.
- I propose to look at this interaction through Imre Lakatos' philosophy of science.
- I suggest that some Lakatos-inspired criteria can help build a model of theory-appraisal in consciousness science.

Abstract

The neuroscience of consciousness is undergoing a significant empirical acceleration thanks to several adversarial collaborations that intend to test different predictions of rival theories of consciousness. In this context, it is important to pair consciousness science with confirmation theory, the philosophical discipline that explores the interaction between evidence and hypotheses, in order to understand how exactly, and to what extent, specific experiments are challenging or validating theories of consciousness.

In this paper, I examine this intricate relationship by adopting a Lakatosian lens and propose that Lakatos' philosophy of science can aid consciousness scientists to better interpret adversarial collaborations in consciousness science and, more generally, to develop a confirmation-theoretic model of theory appraisal in this field.

I do so by suggesting that such a model be built upon three Lakatos-inspired criteria for assessing the relationship between empirical evidence and theoretical predictions: a) the model should represent the *distinction between prediction and accommodation*; b) the model should represent the *structural relevance* of predictions; c) the model should represent the *boldness* of the predictions. I argue that a Lakatosian model of theory-appraisal has both normative and descriptive virtues, and can move the debate forward by acknowledging that theory-appraisal needs to consider the diachronic development of theories, their logical structure, and their relationship with background beliefs and knowledge.

Word count: 8221 (main text); 214 (abstract); 2171 (references).

Introduction

This paper presents some philosophical insights on how to think about the (dis)confirmatory relationship between theories of consciousness and empirical evidence. Thinking seriously about the nature of this relationship is important because the neuroscience of consciousness is seeing a proliferation of theories (Del Pin, Skóra, Sandberg, Overgaard, & Wierzchoń, 2021; Seth & Bayne, 2022; Signorelli, Szczotka, & Prentner, 2021), and consensus is far from near (Francken et al., 2022; Yaron, Melloni, Pitts, & Mudrik, 2022).

Adversarial collaborations, which intend to test different theories of consciousness based on contrasting predictions (Melloni et al., 2023; Melloni, Mudrik, Pitts, & Koch, 2021), promise to reduce the theory-space, and ultimately illuminate the details of how consciousness and brain activity relate. This is surely a welcomed empirical acceleration.

I argue that consciousness science would benefit from directly engaging with confirmation theory, the philosophical field studying the relationship between evidence and hypotheses (Chalmers, 2013; Crupi, 2021; Godfrey-Smith, 2003; Hesse, 1974), since conclusions about the epistemic solidity of theories of consciousness can be unwarranted if not paired with carefully constructed arguments that are sensitive to the nature of the relationship between evidence and theories.

Given that the urge of thinking seriously about confirmation theory in the context of consciousness science is motivated mainly by the goals and the results of adversarial collaborations, I will frame most of the discussion around the first experiment of this sort, designed and performed by the Cogitate Consortium (Cogitate et al., 2023), but my general intention is to provide some suggestions on how to navigate the broader debate on whether, and to what extent, empirical evidence can corroborate or disconfirm theories of consciousness. I will do so by building on Imre Lakatos' view of scientific progress (Lakatos, 1976; Lakatos & Musgrave, 1970). I argue that some Lakatos-inspired criteria can help consciousness scientists build a confirmation-theoretic model of normative theory appraisal for consciousness science. Despite this project being mainly philosophical, I believe it can be seen as complementary to formal approaches that intend to score through probability theory the degree of confirmation that empirical evidence provides to theories (Corcoran, Hohwy, & Friston, 2023).

This project has both descriptive and normative components: it is primarily normative, insofar as it prescribes a way to interpret adversarial collaborations and to guide consciousness scientists towards specific philosophical problems that ought to be solved when building a model of theory-appraisal in consciousness science; and it is descriptive because the Lakatosian lens I apply to consciousness science is already implicit in what consciousness scientists do in practice. A comprehensive and systematic analysis of such practice is nevertheless missing, and that is the descriptive goal of this paper.

The aspiration of this paper is to start a conversation, not to settle it. It is possible that Lakatos' philosophy of science is not the best way to interpret the evidential fit between theories of consciousness and experimental results, but I will build a case that a Lakatosian framework does seem to satisfactorily address the main concerns we should consider when trying to flesh out a confirmation-theoretic framework for consciousness science.

In the first section, I offer a potential reading of the philosophy of science behind theory testing in consciousness science, but I then argue that such a reading is problematic. In the second section, I suggest that we look at empirical theory testing through a Lakatosian lens, and I apply such a lens to consciousness science. In the third section, I specify three criteria that a Lakatos-inspired model of theory-appraisal for consciousness science should consider. In the Discussion, I evaluate the significance of this Lakatosian framework for empirical theory testing in consciousness science. A brief conclusion ends the paper.

1. *Instantia crucis* and severe tests in consciousness science

In the *Novum Organum* (1878), Francis Bacon introduced the notion of *instantia crucis* to refer to a situation able to determine the truth of a hypothesis and, at the same time, the falsity of alternative ones. Philosophers of science (Chalmers, 2013; Crupi, 2021) have seen in this approach the precursor of experimentalism (Hacking, 1982, 1988; Mayo, 1991, 1996), the view that the growth of scientific knowledge is driven by experiments, rather than theories, since scientists can design experiments to bring about a situation of this sort – the *experimentum crucis*.

In this view, scientific progress depends on eliminating predictions that do not conform well to evidence gathered through adequate and severe tests: The growth of scientific knowledge passes through an experimental situation that is able to *eliminate* predictions derived from hypotheses under test. We can call this view “experimental eliminativism”.

An attractive idea (which I will later reject) is to think that this philosophy of science fits nicely with adversarial collaborations in consciousness science. An adversarial collaboration’s goal is to find convergence between proponents of different theories on how to design an experiment that could show a state of affairs that agrees with the prediction of one theory while disagreeing with the prediction of the rival (Clark, Costello, Mitchell, & Tetlock, 2022). Thus, first we identify, with the help of proponents of rival theories of consciousness, an experiment in which leading theories of consciousness generate contrasting predictions, then the experiment will decide which theory conforms better with the observed data (Melloni et al., 2023).

This seems to suggest that an assumption behind adversarial collaborations in consciousness science is that scientific progress passes through accumulation of experimental practice, in line with the main tenet of experimentalism in philosophy of science.

If we were to build a confirmation-theoretic model of theory appraisal based on experimental eliminativism, adversarial collaborations would be seen as an "elimination race": a theory is confirmed if it passes a severe test, and disconfirmed if it does not. Every experiment would correspond to a lap of the race, at the end of which one competitor is eliminated.

Despite some passages seem to suggest this view (for example, Melloni and colleagues write that the adversarial collaboration between two theories "should yield reliable results that can provide substantial evidence for one or the other theory, in order to arbitrate between them" (Melloni et al., 2023, p. 2)), I believe this interpretation would be inappropriate for both normative and descriptive reasons. Although in the literature adversarial collaborations have not been explicitly aligned with experimental eliminativism, this interpretation is sometimes implicit in how adversarial collaboration projects are presented in popular outlets (Finkel, 2023; Lenharo, 2023), and therefore my discussion aims at preventing such interpretation.

Let us focus first on the descriptive reasons for not aligning adversarial collaborations with experimental eliminativism. Take the specific collaboration testing different predictions from the Integrated Information Theory (IIT) (Albantakis et al., 2022; Tononi, Boly, Massimini, & Koch, 2016) and the Global Neuronal Workspace Theory (GNWT) (Dehaene & Changeux, 2011; Mashour, Roelfsema, Changeux, & Dehaene, 2020), led by the Cogitate Consortium. According to IIT, consciousness corresponds to the amount of irreducible causation that a physical system can potentially exert upon itself, which can be mathematically measured through a formalism called Φ. In contrast, for GNWT, consciousness is a property of a representation: a representation becomes conscious when it is broadcast into a "global workspace" in which it becomes available to a host of consuming systems.

Members of the Cogitate Consortium have realized early on that adversaries do not concede easily the elimination of their theories, and rather try to adjust their view in light of the evidence (if they accept the evidence to begin with). This has led members of the Cogitate Consortium to align adversarial collaborations with a sophisticated form of falsificationism (Melloni, 2022) by stating, for example, that the purpose of their adversarial collaboration was "to falsify divergent predictions of IIT and GNWT and not to provide confirmatory evidence" (Cogitate et al., 2023, p. 25), and that it was aimed at "providing the means to change one's mind given contradictory results" (Cogitate et al., 2023, p. 24).

Apart from these descriptive reasons, there are also normative reasons for not aligning adversarial collaborations in consciousness science with experimental eliminativism. This is because not every experiment in this context constitutes a *severe test*, in the eliminativist senses. To see why, we need to focus first on what a severe test is, and then on the specifics of the collaboration led by the Cogitate Consortium.

The most sophisticated version of experimental eliminativism has been arguably developed by Deborah Mayo (Mayo, 1991). According to Mayo, a hypothesis passes a severe test if the (objective) probability of obtaining a specific experimental result is very high if that hypothesis is true and, crucially, very low if the hypothesis is false.

Imagine that we are interested in testing Snell's law of refraction of light through refractive media. We can do so by measuring the angles of incidence and refraction and see whether they coincide with what Snell's law predicts. Assume that they do, but our measurements come with quite large margins of error. In fact, it might happen that alternative theories of refraction (like Ptolemy's theory) predict results that fall within those margins of error, and therefore could account for the experimental result. This means that the probability of obtaining that particular experimental result is quite high even if Snell's law is false, because predictions of alternative theories are not eliminated by this specific measurement. This is why, according to Mayo's view, the hypothesis (i.e., Snell's law) is *not* confirmed by this experiment: it does not pass a severe test.

Vice versa, the famous experiment led by Eddington and Dyson in 1919 (Dyson, Eddington, & Davidson, 1920) for measuring the deflection of starlight in proximity to the sun corroborated Einstein's theory of gravity because it was a severe test for that theory: Einstein's prediction was that the light would deflect at the limb of the sun of 1.75 arc seconds, while the Newtonian theory predicted a deflection of 0.87 arc seconds. The two expeditions reported a deflection of 1.98 ± 0.18 arc seconds and 1.61 ± 0.45 arc seconds, so this was enough to eliminate the predictions of the Newtonian theory of gravity. In this case, then, we have a severe test because the probability of obtaining the results observed by Eddington would be quite low, had Einstein's theory been false.

Do the experiments led by Cogitate have the same degree of "severity"? The Cogitate Consortium have tested these two theories based on three different predictions, but I will now focus mainly on the first. The first prediction is based on different brain areas involved in conscious perceptions: according to IIT, it should be possible to decode stimulus category and orientation, when that stimulus is consciously perceived, from posterior areas only (Boly et al., 2017), while for GNWT it should be necessary to include areas of the prefrontal cortex. (Cogitate et al., 2023) have found that decoding of a consciously perceived stimulus is in fact maximal when based on posterior areas, as predicted by IIT.

Now, if we take experimental eliminativism (at least in Mayo's sophisticated version) as the philosophy of science providing a confirmation-theoretic foundation for the science of consciousness, we would have to admit that this part of the first Cogitate experiment, by itself, is not quite a severe test. IIT and GNWT are not the only options available in the theoretical landscape of consciousness science, and several competing theories of consciousness also predict that frontal areas are necessary for conscious perception (Brown, Lau, & LeDoux, 2019; Graziano & Webb, 2015; Lau, 2022), while some others predict that posterior areas are sufficient (Lamme, 2006, 2010). This means that it is not true that there would be low probability of maximally decoding conscious contents, if GNWT were false (since one could find frontal decoding due to another mechanism that is not related to a global workspace – e.g., a higher order representation); and moreover, it is not the case that if IIT were false, then the probability of decoding conscious contents from posterior areas only would be low (since maximal decoding could depend on local recurrent activity). Therefore, Mayo’s severity criterion is not satisfied. But in this view, if an experiment is not a severe test, then it cannot contribute much to the growth of scientific knowledge.

My claim is that this conclusion, if applied to the series of experiments performed by the Cogitate Consortium, is unacceptable, because the experiments do challenge both theories, and push theorists to either dismiss the evidence for methodological limits (however, this option should be prevented by the very nature of adversarial collaborations) or adjust their theories in light of empirical evidence. Thus, the Cogitate experiment contributes to substantial progress in consciousness science despite not being a severe test in the experimental eliminativist sense.

There is a further problem with experimental eliminativism, which can be seen by analysing the other two predictions tested by the first Cogitate experiment: The second prediction was based on whether the neural activity that correlates with conscious contents is sustained for the duration of the experience (as predicted by IIT) or, instead, phasic, with spikes of activity correlating with stimulus onset and offset (as predicted by GNWT). The third prediction focused instead on interareal connectivity corresponding to conscious perception, with IIT predicting high interconnectivity between posterior regions for the duration of the experience, and GNWT predicting phasic connectivity between category selective areas and the prefrontal cortex.

The results here were mixed, with the second prediction favouring IIT, and the third prediction being more aligned with GNWT.

The crucial issue with experimental eliminativism is that it does not clearly specify how to connect these contrasting pieces of evidence with high-level theories. Here, I build on the idea that hypotheses can be formulated at various levels of analysis: they can be extremely specific and fine-grained, or they can be more general, less detailed, and function as working hypotheses. Or, they can be formulated at an even more general level as research programmes, theories, or paradigms (Douglas & Magnus, 2013).

The alleged “elimination race” posited by experimental eliminativism would be between low-level, very fine-grained, hypotheses (e.g., whether the hypothesis that neural activity during conscious perception is phasic should be rejected), but it is not clear whether, or at which point, the high-level theory generating that prediction (e.g., GNWT) should be rejected. Indeed, it is not clear whether *theories* of consciousness could be part of the race at all: This view is silent on how low-level experimental evidence connects to high-level theories (for criticisms of Mayo’s view along these lines, see (Douglas & Magnus, 2013; Musgrave & Mayo, 2009; Worrall & Mayo, 2009)). Therefore, experimental eliminativism is unable to provide a picture of how scattered and sparse low-level evidence can inform our normative appraisal of theories of consciousness.

This discussion suggests that interpreting the philosophy of science at the basis of adversarial collaborations as an experimental eliminativism would be inappropriate, for both descriptive and normative reasons.

What we need is a philosophical view of scientific practice able to account for: i) the informativeness of experiments carried out in the context of adversarial collaborations; and ii) a unified picture of high-level theories and their low-level experimental predictions.

I believe that such a framework can be provided by applying the philosophy of science of Imre Lakatos to consciousness science. A Lakatosian framework, I contend, could also help reframe our expectations from theory testing in consciousness science: we could perhaps overcome any “elimination” talk, and interpret instead experimental work as an opportunity for theoretical self-improvement (in the context of adversarial collaborations, this has been noticed by (Cowan et al., 2020) and is very much in line with the sophisticated falsificationist attitude of (Melloni, 2022) and (Cogitate et al., 2023)).

2. Lakatos and the science of consciousness

In this section, I will briefly present the main ideas of Lakatos’ philosophy of science (Lakatos, 1968a, 1968b, 1974, 1976; Lakatos & Musgrave, 1970), and I will then apply them to consciousness science.

The Lakatosian view of science builds on the idea that scientific theories have two components: the *core*, namely the claims that characterize the essence of the theory, and the *belts of peripheral auxiliary hypotheses.* The conjunction of core theses and auxiliary hypotheses entails the theory's predictions, and experimental testing targets these predictions.

Lakatos agreed with Popper (Popper, 1959) that scientific progress is not made by verifying predictions entailed by theories, but by trying to falsify them. However, if a theory is made of a conjunction of core and auxiliaries, a falsified prediction does not automatically disconfirm the core claims of the theory, but rather it disconfirms the conjunction of core and auxiliaries, and a conjunction is false if only one of the conjuncts is false. This means that scientists can interpret the experimental falsification of a prediction by holding on to the core claims of their theory, and by rejecting some of the auxiliaries. This makes theory testing an arduous affair, since, as Duhem and Quine realised, it becomes difficult to disentangle the theoretical aspects put under pressure by empirical testing from all the background beliefs held by scientists (Duhem, 1954; Quine, 1951).

Lakatos' view, which was substantially informed by the history of science, is that scientists generally opt for revising some of the auxiliary hypotheses while holding on to the core claims of the theory. This generates a *research programme,* a diachronically constituted scientific effort built around a core set of theses, assumptions, and heuristics: within a research programme, the core remains the same, while the belts of peripheral auxiliary hypotheses are modified under the pressure of empirical testing.

Now, scientists can modify the theory by changing its empirical or theoretical content. The *theoretical* content of the research programme requires that the new theory generates novel testable predictions, while the *empirical* content requires that at least some of these novel predictions be true. If a research programme fails in at least one of these two aspects, it can be said to be *degenerating*, rather than *progressive*, and therefore it does not constitute good science (for an historical case, see (Scerri & Worrall, 2001)). However, for Lakatos, a degenerating research programme can ultimately be discarded only by *another research programme*, not by an experiment (Lakatos, 1974).

This brief presentation of Lakatos' view should suffice to see why a Lakatosian lens on consciousness science can solve the problems that emerged by interpreting theory testing through the experimental eliminativist lens.

First, experiments like the one designed and performed by the Cogitate Consortium are informative and contribute substantially to the growth of knowledge in consciousness science, insofar as they provide a falsification-driven test for the empirical content of consciousness research programmes and constrain the possible moves theorists can make to adjust their theories. This is actually hinted at by the members of the Consortium, who write that the adversarial collaboration was aimed at "providing the means to change one's mind given contradictory results" (Cogitate et al., 2023, p. 24). Experiments are thus seen as informative as long as they have the potential to disrupt the empirical content of the theory and therefore generate a research programme that, in the longer run, can turn out to be progressive or degenerating. The Lakatosian framework seems to capture quite well the intent of consciousness scientists actively involved in this adversarial collaboration, which is a positive descriptive virtue of the framework.

Second, the Lakatosian framework admits that scientists can in fact be rational when they hold on to their theory even when presented with some contrasting evidence. Again, this is because a refuted prediction can simply signal that the specific connection that the scientist drew between core claims and auxiliary hypotheses is incorrect; it does not necessarily mean that the theoretical core is wrong. This is descriptively accurate too: for example, in (Cogitate et al., 2023, p. 27), Dehaene defends GNWT in light of some contrasting evidence, and claims that a proper evaluation of the theories should wait for the second Cogitate experiment (Melloni et al., 2023). And normatively speaking, if the Lakatosian framework is on the right track, he is entirely rational in doing so.

Third, the connection between theory and prediction is a natural consequence of the distinction between core and protective belts of auxiliaries (or between centre and periphery): predictions are (deductively) derived from the conjunction of core and auxiliary hypotheses, which means that, as Musgrave (2023) notes, auxiliary hypotheses have two jobs in the Lakatosian picture: on the one hand, they protect the core claims from *direct* falsification; on the other hand, they connect theoretical claims with experimental practice. A theory will eventually be refuted when its core will be unable to parsimoniously relate to empirical evidence and background knowledge, and will be superseded in that regard by a rival theory: the normative appraisal of research programmes depends on how well low-level predictions do experimentally, *and* on how well the scientists will be able to modify the theory in light of experimental evidence.

The Lakatosian framework is thus quite promising in interpreting adversarial collaborations in consciousness science and can shed light on the general nature between evidence and hypotheses. I believe that this framework is conducive to the claim that adversarial collaborations in consciousness science should be seen as theoretical self-improvement rather than as an “exclusion race". This, in turn, can inform a more general confirmation-theoretic view on how experimental evidence corroborates high-level theories of consciousness.

In order to use this framework to build a model of theory-appraisal for consciousness science it is crucial to specify the Lakatos-inspired criteria upon which this model could/should be built. These criteria are: a) the model should represent the *distinction between prediction and accommodation*; b) the model should represent the *structural relevance* of predictions; c) the model should represent the *boldness* of the predictions.

3. Three Lakatosian criteria for theory-appraisal

In this section, I briefly introduce the criteria, inspired by the Lakatosian view of science, that can be used to build a model of theory-appraisal for the neuroscience of consciousness. These criteria are probably not exhaustive and are sketched here at a general level so as to point at the type of philosophical considerations that could be beneficial to consciousness scientists interested in building a model of theory-appraisal. The specific approach one takes with respect to these criteria is then a modeller’s decision, and it is possible that different formal models can be built by following the same set of criteria. My goal here is to just flag that whatever choice the modeller makes, it will probably come with substantial philosophical assumptions, and so I will make clear what these assumptions amount to.

It is important to clarify that these confirmation-theoretic criteria are not all explicitly stated by Lakatos but are rather extrapolated from his overall view.

3.1. Prediction vs Accommodation

Philosophers of science tend to distinguish predictions from accommodations. To use Nozick's vivid example, imagine two archers: one hits the bullseye, while the other hits a random white wall and then draws a bullseye around the arrow (Nozick, 1983, p. 9). The first would be like a predictor, the second like an accommodator.

Predictivism is the philosophical view that predictions bear higher confirmatory power than accommodations (Douglas, 2009; Douglas & Magnus, 2013; Lipton, 1990; Maher, 1988; Worrall, 1989), and it seems to be at the core of adversarial collaborations in consciousness science because theories are not supposed to be tested through pieces of evidence that they can accommodate, but by evidence they can predict.

The distinction is particularly relevant for a Lakatosian framework: as seen in the previous section, progressive research programmes are able to predict novel facts, while degenerating programmes fail to do so. The notion of "novel" fact is thus crucial to evaluate the progressivity of a research programme, and since progressivity is based on predictions, rather than accommodations, the distinction between prediction and accommodation rests on what "novel" means.

At first glance, it might not seem troubling: a prediction is about a state of affairs the occurrence of which is unknown at the time of the formulation of the prediction, while an accommodation occurs when the state of affairs is already known at the time a theory is formulated, and the theory is able to account for it. This is the *temporal* reading of the prediction/accommodation distinction, based on the view that novel facts are temporally novel, and is arguably what (Popper, 1968) and Lakatos (Lakatos, 1968b) had in mind.

However, the case of Mercury's perihelion, an anomaly in the Newtonian theory of gravity, poses a problem for this temporal view of "novel facts". The anomaly of Mercury's perihelion was well known *before* Einstein's theory of gravity came along; and yet, Einstein included the measurement of the precession of Mercury's perihelion as one of the three key tests for general relativity, because if the theory was right, the curvature of spacetime would have naturally accounted for the precession (Einstein, 1916). Einstein's view was that it would be wrong to think that the precession of Mercury's perihelion does not have much confirmatory power with respect to general relativity: the anomalous behavior of Mercury's perihelion just naturally follows from general relativity, and therefore it confirms the theory even if it was already known when the theory was formulated.

Lakatos, following Zahar (Lakatos & Zahar, 1975; Zahar, 1973), accepted Einstein's suggestion that novelty is not a temporal notion, and modified his theory by claiming that a fact is novel insofar as it was not part of the empirical facts that motivated the construction of the theory in the first place. Einstein did not build general relativity to explain Mercury's perihelion, and that is why general relativity does not accommodate the precession of Mercury's perihelion. Rather, it (successfully) predicts it. This is what philosophers of science call "use-novelty" or "heuristic-novelty" – for a similar view, see (Worrall, 1985, 1989).

This debate can be translated into consciousness science. Consider the results of Sperling's partial report paradigm (Sperling, 1960), where Sperling demonstrated that, even if subjects can normally report four or five items out of a larger array of letters, organised in a 3 by 3 or 4 by 4 matrix, they are nonetheless able to report almost all the letters in a row if a row is cued immediately after the array of letters disappears. This happens independently of the cued row, suggesting that subjects are conscious of more than what they can normally report ((Block, 2011) – for an alternative interpretation, see (Phillips, 2011)).

IIT was not built to account for the results obtained in Sperling's partial report experiments, but such results can be accounted for by IIT quite naturally (Haun, Tononi, Koch, & Tsuchiya, 2017; Tononi, 2015; Tononi et al., 2016, pp. 456-457). In fact, Tononi (2015) writes that

> IIT emphasizes that the set of elements (neurons) that specify a particular concept within a conceptual structure may be more or less difficult to access from within the complex [...]. The concepts that can be accessed and communicated to an external observer at any given time are a minimal subset of the entire set of concepts that compose the quale sensu lato [...] nor can one easily communicate the relationships among them (distance in cause-effect space) that give each experience its particular meaning.[1]

This suggests that IIT predicts that phenomenal experience (i.e., the overall cause-effect structure composed of sub-structures and relations among them) goes beyond what can be reported (Block, 2011), and can thus account for Sperling's results.

So, Sperling's experimental results count as a "use-novel" fact for IIT, even if the Sperling experiment predates the development of IIT by about four decades. But take now Lamme's recurrent processing theory (RPT) (Lamme, 2010), the view that feedback loops within sensory areas are necessary and sufficient for sensory consciousness. Lamme (2006, 2010) states that Sperling's findings are among the motivating grounds for his theory, and this means that Sperling's experimental results are accommodated, and not predicted, by RPT.

The opposite is true for the evidence that the cerebellum can be removed without impacting consciousness. Even if the foundation of IIT is phenomenological (Ellia et al., 2021; Oizumi, Albantakis, & Tononi, 2014), the fact that the cerebellum is not necessary for consciousness is often presented by IIT proponents as one of the facts that a theory of consciousness needs to explain, and is therefore part of the empirical motivating ground for the development of IIT (Massimini & Tononi, 2018). If this is right, then the fact that the cerebellum is not necessary for consciousness is accommodated by IIT, but it counts as a "use-novel" fact for RPT, because RPT can easily explain this fact by pointing out that the neuroanatomical connectivity of the cerebellum is mostly based on

[1] In the current version of the theory (Albantakis et al., 2022), "concepts" are called "phenomenal distinctions" and conceptual structures are "cause-effect structures", or "Φ-structures".

feedforward connections. In this case, then, the fact that the cerebellum is not necessary for consciousness is confirmatory for RPT, but not (or only partially, if accommodations have some confirmatory value) for IIT.

The present discussion indicates that a general model of theory-appraisal should be sensitive not only to the logical relation between evidence and hypotheses, but also to how, and in which context, the hypothesis is formulated. The reason why this discussion is relevant is that if an experiment (either based on adversarial collaboration or not) tests only some theories of consciousness at a time, it might be possible that a correct prediction of a theory "T" could be easily accounted for by an alternative theory "T1" that is not directly tested in that specific experiment. That would count as a novel correct prediction (both in temporal and "use-novelty" sense) for T, but also as a use-novel correct prediction for T1: experiments could have confirmatory value for theories that are not directly tested in that particular setting.

There is a further complication, though. Consider Mashour et al.'s claim (Mashour et al., 2020, p. 787) that GNWT can easily explain the finding, supposedly supporting IIT, that the reduction in complexity of brain activity can reliably predict whether an unresponsive subject is dreaming, in a state of unresponsive wakefulness, or non-conscious (Casali et al., 2013; Massimini, Boly, Casali, Rosanova, & Tononi, 2009). Mashour et al. write that "Although this experiment was inspired by an alternative theory of consciousness, [...] the results are fully compatible with the GNW, which predicts that the conscious state leads to a deeper and more prolonged propagation of activation through long-distance connections compared to the unconscious state" (Mashour et al., 2020, p. 787). Although there might be issues with how to formally interpret the notion of brain complexity, and therefore on whether there is convergence between the formalism that is supposed to support IIT and its compatibility with GNWT (Farisco & Changeux, 2023; Sarasso et al., 2021), we can accept for the sake of the argument that these experimental findings do count as a use-novel fact for GNWT. If that is the case, they should be confirmatory relevant for GNWT under a "use-novel" framework.

However, a Lakatosian scholar like Nunan (1984) has argued that a fact can count as novel only if it is not predicted by an alternative research program, on the basis that such a novelty criterion is the only one that is able to connect the progressivity of a research programme with rival programmes, and therefore able to demonstrate why the progressive research programme should be rationally chosen over the rivals (see also (Lakatos, 1974)). In a similar spirit, Douglas and Magnus (2013) write that "new evidence or new evidential relations throw an evidential gauntlet down for competitor theories — they must *accommodate* the new evidence or risk epistemic demerit" (Douglas & Magnus, 2013, p. 13; emphasis added). If this is right, GNWT would *not* predict the experimental finding of Casali et al. (2013), not even in a "use-novel" sense – it would merely accommodate it, and therefore that empirical evidence could not contribute much to confirming GNWT.

Here, my goal is not to defend a specific way of drawing the distinction between prediction and accommodation, but to flag some crucial questions that are preliminary to any model of theory-appraisal that can be employed in evaluating and comparing theories of consciousness. It is important that a model of theory-appraisal be sensitive to these concerns, and it will ultimately be a theoretical (and philosophical) choice of the modeller how to specifically interpret the novelty of a prediction, and how to value the confirmatory aspect of accommodations.

3.2. The structural relevance of predictions

The second criterion builds upon the Lakatosian distinction between the core theses of a theory and its auxiliary hypotheses. This is relevant to theory testing in consciousness science, because some have remarked that the first adversarial collaboration testing contrasting predictions from IIT and GNWT did not test the theories directly. For example, Dehaene writes that "none of the massive mathematical backbone of IIT, such as the φ measure of awareness, was tested in the present experiment" (Cogitate et al., 2023, p. 27) – see also (Fleming, 2023). However, a similar point could be raised for GNWT too, as correctly pointed out by (Seth, 2023).

If we take the Lakatosian picture seriously, this worry is not a concern *per se*, but just the rule of theory testing: scientific theories are not tested at their core, but rather through the belts (which are called "protective" precisely for this reason) of auxiliary hypotheses.

This is also the lesson learnt from the Duhem-Quine thesis: given the immense web of beliefs held by scientists, it is always possible to fit the data with any theory by adjusting and modifying the related auxiliaries, and therefore a *direct* empirical test of theoretical claims is fundamentally impossible.

My claim here is that we can circumvent this concern by noticing that *the degree of confirmation a prediction confers to a theory partly depends on how peripheral the prediction is*: the farther away from the core the prediction is, the less (dis)confirmatory power will bear.

To substantiate this discussion, let us represent (at a general level) the structure of IIT and GNWT through a Lakatosian lens. The basic idea is that each level of the periphery generates predictions based on what is assumed and predicted by the previous level, with more peripheral levels specifying at a finer grain the predictions of the levels closer to the centre, thus making the theory more detailed (for a similar view, see (Henderson, Goodman, Tenenbaum, & Woodward, 2010)).

For IIT, the core claims are its phenomenological axioms (which define the essential properties of consciousness), its postulates (which state how those properties can be accounted for in physical terms), and the explanatory identity stating that consciousness is integrated information (Albantakis et al., 2022). Outside of the core we can find the mathematical translation of the postulates, and the claim that consciousness can be measured through the *specific* formal apparatus of the theory. Notice that the *general level* claim that consciousness can be quantitatively measured through integrated information is at the core of IIT, but the specific claim that consciousness is identical to the specific formalism presented in (Albantakis et al., 2022) is outside of the core: the specific formalism to determine Φ has changed throughout the various iterations of the theory (Barbosa, Marshall, Albantakis, & Tononi, 2021; Oizumi et al., 2014; Tononi, 2004, 2012), while the general idea that consciousness is integrated information has remained constant.

In IIT, then, the first belt of auxiliaries just requires background knowledge of mathematical nature to derive the specific Φ-measure from IIT's core.

From this first belt of auxiliary hypotheses, we can move to the next, and the distance we travel from the core will depend mainly on the number of auxiliary background assumptions we need to add to core claims in order to derive a specific prediction. For example, IIT's prediction that the back of the brain should be sufficient for conscious perception depends on *neuroanatomical* background knowledge telling us that posterior cortical areas, and not the prefrontal cortex, are constituted by neuronal grids that have the appropriate physical structure to sustain large Φ values (Grasso, Haun, & Tononi, 2021). Thus, to derive the prediction that the back of the brain is sufficient for conscious perception we need the specific mathematical formalism specified by the first peripheral belt *and* some neuroanatomical background knowledge. In this way, we move from the first belt of auxiliary hypotheses to the second belt.

A similar analysis can be done with respect to GNWT: Its core claim is of cognitive nature and states that conscious perception occurs when information processed by modular and local processors is broadcast into a global workspace which renders it available to various consumer systems (Baars, 1988). This core claim is then made more precise by postulating that the neuronal implementation of the global workspace requires pyramidal neurons with long-range axons (Dehaene, Kerszberg, & Changeux, 1998), which requires auxiliaries of neurophysiological nature. The theory can then predict that the prefrontal cortex is necessary for consciousness because it exhibits "the greater density of neurons thought to be critical for global broadcasting of information" (Mashour et al., 2020, p. 777), and this prediction requires neuroanatomical auxiliaries (for in-depth discussions of the structure and historical development of the global workspace hypothesis, see (Mashour et al., 2020, p. 776) and (Baars, Geld, & Kozma, 2021)).

This discussion matters for adversarial collaborations because it is possible that a successful prediction might corroborate one theory while disconfirming the other, but the extent to which it does so might depend on how close that prediction is to the core of each theory, and possibly on the nature of the prediction, given the nature of the core (e.g., computational background assumptions might be more relevant than assumptions on the mechanistic details of the brain, if the core of the theory is computational – see (Fleming, 2020; Lau, 2022; Lau & Rosenthal, 2011; Wiese & Friston, 2021)): the idea is that the relevance of a prediction is associated to the disruptive power the prediction bears, and the closer the prediction is to the core, the higher the power (compare: the Mayor of London is probably aware that closing Liverpool Street Station causes much more troubles to their city than closing Cockfosters Station).

Imagine an experiment involving a prediction that makes use of certain assumptions on the link between attention and consciousness. The result of that experiment will be much more relevant for theories that see consciousness and attention closely related (Graziano, 2022; Parr, Corcoran, Friston, & Hohwy, 2019; Vilas, Auksztulewicz, & Melloni, 2022), but only partially relevant for theories like IIT, that are not built around a connection between consciousness and cognition, and therefore require a large number of background assumptions to formulate predictions about the relation between consciousness and specific cognitive phenomena like attention. In case of contradictory evidence, IIT would just need to revise some auxiliaries about the relationship between integrated information and attentional processes, which would not significantly impact its core claims, and for this reason, although the experiment would still be informative, the disruptive power of the prediction would be limited.

The criterion of structural relevance thus maintains that a model of theory-appraisal should consider the distance between core claims and the peripheral level at which the experimental prediction is formulated: the longer the distance, the less relevant the prediction in confirming or disconfirming a theory.

3.3. The boldness of predictions

The third Lakatosian criterion for a model of theory-appraisal stems from the Popperian tenet that good scientific theories should exhibit an element of risk: the criterion states that the confirmatory power of a prediction is partly constituted by its *boldness*, where boldness is intended as a measure of divergence from consensus and background knowledge. A bold prediction is a prediction that is peculiar to one theory, that sets it apart from competitors, and it would be very unlikely to be verified if the theory were false[2]. Predictions are bold when they are risky (i.e., depart from consensus) and meaningful (i.e., they single out a specific theory)[3]. As Lipton puts it, "evidence that discriminates between competing theories is more valuable than evidence that is compatible with all of them" (Lipton, 1990, p. 54).

There is a sense in which this criterion aligns with Mayo's intuition that the confirmatory status of a prediction depends on a test that minimizes the chance that the prediction could turn out to be correct even if the theory is false: we want a prediction to be confirmatory of a specific theory because it is entailed by that theory and *that theory alone.*

For this reason, boldness goes hand in hand with the idea of *precision*, although the two concepts are not exactly equivalent. The idea is that a prediction is precise when it limits vagueness, and it is clear in discriminating between prohibited states of affairs – the larger and better defined the set of prohibitions, the more precise the prediction is. For example, predicting that tomorrow will be sunny and windy is more precise than predicting that tomorrow will be sunny, but it might not be a bold prediction, if several different meteorological models predict that tomorrow will be sunny and windy. Thus, boldness discriminates between *theories*, given the predicted evidence, while precision discriminates between *states of affairs*. Despite this difference, the two concepts are related because it is easier for an imprecise prediction to be accounted for by a larger set of theories, since imprecise predictions rule out fewer states of affairs than precise ones (for the idea that imprecise predictions have epistemic value nonetheless, see (Elliott-Graves, 2020)). Here, I will formulate the boldness criterion assuming that bold predictions must also be precise predictions.

[2] There is evidence from psychological research that people value bold hypotheses more than relatively uninformative ones (McKenzie & Amin, 2002). This psychological fact can be nicely captured by the Lakatosian model I am suggesting, which speaks in favour of the model.

[3] Special thanks to Liad Mudrik for pointing out this distinction to me.

To see how the criterion works in practice, consider another prediction of IIT. This is the quite surprising prediction that a network of inactive (but not deactivated) neurons can contribute to consciousness (for a critical discussion that questions whether this prediction is testable at all, see (Bartlett, 2022)). The prediction derives from core claims of IIT, like the idea that what matters for consciousness is the intrinsic causal *powers* (i.e., how neurons *can* influence other neurons, rather than *whether* they do so) of a physical system, together with auxiliary assumptions that bridge the general-level claims in the theory's core to the neurophysiological context, and specifically the brain's metabolic constraints (Balduzzi & Tononi, 2009, p. 15).

This is a bold prediction of IIT, in the sense that the idea that inactive neurons can sustain consciousness as much as active ones is a *unicum* among neuroscientific theories of consciousness, and therefore sets IIT apart from neuroscientific consensus. Tononi and Koch are aware of this, and write that "A theory is the more powerful the more it makes correct predictions that violate prior expectations" (Tononi & Koch, 2015, p. 9).

A final point on the relation between boldness and structural relevance: particular predictions that set apart a theory in the theory-space, and therefore score high in boldness, should be expected to be quite close to the core, and therefore score high in structural relevance. This is because the structural relevance of a prediction partly depends on the number of auxiliary background assumptions used to derive that prediction, and it is easier to formulate a hypothesis that substantially diverges from consensus and background knowledge if its construction is influenced more by the 'gravitational' pull of core claims, rather than by large part of more commonly shared background knowledge.

However, the two concepts remain distinct, since structural relevance refers to the informativeness of a prediction with respect to the theory's core, while boldness refers to the informativeness of a prediction with respect to other theories in the theory-space. And given this independence, it might be possible that certain predictions that are quite peripheric turn out to be bolder than certain predictions that are closer to the centre.

4. Discussion

How can these three criteria be practically applied to theory-appraisal in consciousness science? My discussion here has been qualitative, as I focused on the nature of the confirmatory relationship between evidence and theories, but the discussion has been built on the idea that there can be *degrees* of (dis)confirmation. The idea that a prediction comes with an associated value of confirmatory power, and that such value can be mathematically formalised, is generally the hallmark of Bayesian confirmation theory (Howson & Urbach, 1989). It is thus reasonable to think that a formal confirmation-theoretic model of theory-appraisal will be based on some form of Bayesianism and will prescribe that consciousness scientists are rational as long as they modify their prior beliefs according to Bayes' conditionalization rule (Corcoran et al., 2023).

A further question concerns the specific way that the concepts introduced here (e.g., disruptive power, boldness, etc.) can be precisely quantified, which is a topic that could be explored in future work.

As I said above, these Lakatos-inspired criteria do not intend to resolve the discussion about how to model the relationship between evidence and hypotheses in consciousness science, and it might very well be that a Bayesian account could complement and ameliorate the Lakatosian framework presented here (even if a full-blown combination of Lakatos and Bayes might be hard to achieve, see (Lakatos, 1968a)). However, I stress that the importance of building a confirmation-theoretic model of theory-appraisal in consciousness science upon Lakatosian guidelines lies in the emphasis that this framework puts on the structural nature of theories, and on the idea that the normative appraisal of theories does not only depend on the relationship between evidence and hypotheses, but also on the diachronic development of the research programme.

This helps avoid any "elimination" talk (at least in the short term) in regard to adversarial collaborations and empirical testing of theories of consciousness. Rather than a race between competitors, a Lakatosian lens sees empirical testing of theories of consciousness as pushing theorists to revise and modify their theory by readjusting the relationship between core and auxiliaries. This amounts to interpreting theory testing as a sort of empirical 'training ground' for theoretical self-improvement.

Moreover, this lens opens the door to an interpretation of scientific progress based on the social and historical situatedness of scientists. As rightly noted by Michela Massimi, scientific agreement

> is not a matter of winners or losers; of one scientific perspective prevailing over another, or imposing itself on others. It is instead a matter of science being a fundamentally social and cooperative inquiry, where progress takes place not *in spite of* but *thanks to* a plurality of scientific perspectives. Scientific progress is ultimately the story of our coming to agree whilst perspectivally disagreeing (Massimi 2021, p. S6124; italics in the original).

I believe that this is very much in line with how theories of consciousness are tested, especially thanks to adversarial collaborations. In fact, some of the Lakatosian ideas I am suggesting have been more or less explicitly endorsed by scientists involved in adversarial collaborations. In this context, falsifying a prediction does not amount to refuting a theory, in line with Lakatos' sophisticated falsificationism. This speaks in favour of the picture I am proposing, which can be seen as a way to systematize and ground on philosophical foundations what is already done (at least in part) by consciousness scientists.

In sum, the Lakatosian picture sketched here can i) provide a philosophical foundation to interpret the effort of adversarial collaborations in consciousness science by acknowledging the informativeness and relevance of experiments carried out in the context of adversarial collaborations; ii) provide normative guidance on how to better interpret the nature of the relationship between evidence and hypotheses, and thus support the development of a model of theory-appraisal; iii) explain why consciousness scientists are not dismissing their theories in light of some contradictory evidence, and prescribe that they are rational in doing so; and iv) accurately describe the intentions and the practices of scientists who are actively testing theories of consciousness.

Conclusion

In this paper, I have provided a philosophical foundation for the debate of how theories of consciousness fit with empirical evidence. This has been motivated by the objective of setting the stage for the development of a confirmation-theoretic model of theory-appraisal for consciousness science.

I have argued that the Lakatosian framework I am suggesting can help the development of a model of theory-appraisal by pushing modellers to consider three different criteria of how empirical evidence relates to theories: these are i) the distinction between prediction and accommodation; ii) the structural relevance of predictions; and iii) the boldness of predictions.

The picture depicted here sees empirical theory testing as an opportunity for theoretical self-improvement, where scientific progress is ultimately driven by the ability of scientists to collaboratively devise ways to look for contradictory evidence and ameliorate their theories in light of such evidence.

A discussion on the nature of the relationship between evidence and theories is particularly important in a field like consciousness science, where theories have been traditionally built and tested in theoretical silos, and adversarial collaborations are challenging theorists to revise their theories given the evidence. Such an empirical acceleration is surely beneficial, and the present paper intends to sketch a possible way to philosophically complement this empirically driven development of the neuroscience of consciousness and its theories.

Acknowledgments

I wish to thank Jakob Hohwy, Andrew Corcoran, Rony Hirschhorn, Liad Mudrik, Lucia Melloni and Francesco Ellia for their helpful comments on previous versions of this paper, and for many insightful conversations on these topics. This research was funded in part by the ARC TWCF INTREPID grant. I also wish to acknowledge the support of the Azrieli Foundation.

References

Albantakis, L., Barbosa, L. S., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., . . . Tononi, G. (2022). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *arXiv*. doi:10.48550/ARXIV.2212.14787

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*: Cambridge University Press.

Baars, B. J., Geld, N., & Kozma, R. (2021). Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments. *Frontiers in Psychology, 12*. doi:10.3389/fpsyg.2021.749868

Bacon, F. (1878). *Bacon's Novum Organum*. Oxford: Clarendon Press.

Balduzzi, D., & Tononi, G. (2009). Qualia: the geometry of integrated information. *PLoS Comput Biol, 5*(8), e1000462. doi:10.1371/journal.pcbi.1000462

Barbosa, L. S., Marshall, W., Albantakis, L., & Tononi, G. (2021). Mechanism Integrated Information. *Entropy, 23*(3), 362. Retrieved from https://www.mdpi.com/1099-4300/23/3/362

Bartlett, G. (2022). Does integrated information theory make testable predictions about the role of silent neurons in consciousness? *Neuroscience of Consciousness, 2022*(1). doi:10.1093/nc/niac015

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences, 15*(12), 567-575. doi:10.1016/j.tics.2011.11.001

Boly, M., Massimini, M., Tsuchiya, N., Postle, B. R., Koch, C., & Tononi, G. (2017). Are the Neural Correlates of Consciousness in the Front or in the Back of the Cerebral Cortex? Clinical and Neuroimaging Evidence. *J Neurosci, 37*(40), 9603-9613. doi:10.1523/JNEUROSCI.3218-16.2017

Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences, 23*(9), 754-768. doi:10.1016/j.tics.2019.06.009

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., . . . Massimini, M. (2013). A Theoretically Based Index of Consciousness Independent of Sensory Processing and Behavior. *Science Translational Medicine, 5*(198), 198ra105-198ra105. doi:10.1126/scitranslmed.3006294

Chalmers, A. F. (2013). *What Is This Thing Called Science?* : Hackett Publishing Company, Incorporated.

Clark, C. J., Costello, T., Mitchell, G., & Tetlock, P. E. (2022). Keep your enemies close: Adversarial collaborations will improve behavioral science. *Journal of Applied Research in Memory and Cognition, 11*(1), 1-18. doi:10.1037/mac0000004

Cogitate, Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., . . . Melloni, L. (2023). An adversarial collaboration to critically evaluate theories of consciousness. *bioRxiv*, 2023.2006.2023.546249. doi:10.1101/2023.06.23.546249

Corcoran, A. W., Hohwy, J., & Friston, K. (2023). Accelerating scientific progress through Bayesian adversarial collaboration. *SSRN*. doi:Available at SSRN: https://ssrn.com/abstract=4548942 or http://dx.doi.org/10.2139/ssrn.4548942

Cowan, N., Belletier, C., Doherty, J. M., Jaroslawska, A. J., Rhodes, S., Forsberg, A., . . . Logie, R. H. (2020). How Do Scientific Views Change? Notes From an Extended Adversarial Collaboration. *Perspectives on Psychological Science, 15*(4), 1011-1025. doi:10.1177/1745691620906415

Crupi, V. (2021). Confirmation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2021 ed.): Metaphysics Research Lab, Stanford University.

Dehaene, S., & Changeux, J.-P. (2011). Experimental and Theoretical Approaches to Conscious Processing. *Neuron, 70*(2), 200-227. doi:https://doi.org/10.1016/j.neuron.2011.03.018

Dehaene, S., Kerszberg, M., & Changeux, J.-P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A, 95*(24), 14529-14534. doi:10.1073/pnas.95.24.14529

Del Pin, S. H., Skóra, Z., Sandberg, K., Overgaard, M., & Wierzchoń, M. (2021). Comparing theories of consciousness: why it matters and how to do it. *Neuroscience of Consciousness, 2021*(2). doi:10.1093/nc/niab019

Douglas, H. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science, 76*(4), 444-463. doi:10.1086/648111

Douglas, H., & Magnus, P. D. (2013). State of the Field: Why novel prediction matters. *Studies in History and Philosophy of Science Part A, 44*(4), 580-589.

Duhem, P. M. M. (1954). *The Aim and Structure of Physical Theory*: Princeton University Press.

Dyson, F. W., Eddington, A. S., & Davidson, C. (1920). IX. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 220*(571-581), 291-333. doi:doi:10.1098/rsta.1920.0009

Einstein, A. (1916). Die Grundlage der allgemeinen Relativitätstheorie. *Annalen der Physik, 354*(7), 769-822. doi:https://doi.org/10.1002/andp.19163540702

Ellia, F., Hendren, J., Grasso, M., Kozma, C., Mindt, G., P. Lang, J., . . . Tononi, G. (2021). Consciousness and the fallacy of misplaced objectivity. *Neuroscience of Consciousness, 2021*(2). doi:10.1093/nc/niab032

Elliott-Graves, A. (2020). The Value of Imprecise Prediction. *Philosophy Theory and Practice in Biology, 4*(12).

Farisco, M., & Changeux, J.-P. (2023). About the compatibility between the perturbational complexity index and the global neuronal workspace theory of consciousness. *Neuroscience of Consciousness, 2023*(1). doi:10.1093/nc/niad016

Finkel, E. (2023). 'Adversarial' search for neural basis of consciousness yields first results. *Science, 380*(6652). doi:10.1126/science.adj3877

Fleming, S. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness, 2020*(1). doi:10.1093/nc/niz020

Fleming, S. (2023). The state of consciousness science – and why the media got it wrong. Retrieved from https://elusiveself.wordpress.com/2023/07/20/the-state-of-consciousness-science-and-why-the-media-got-it-wrong/

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neuroscience of Consciousness, 2022*(1). doi:10.1093/nc/niac011

Godfrey-Smith, P. (2003). *Theory and Reality: An Introduction to the Philosophy of Science*: University of Chicago Press.

Grasso, M., Haun, A. M., & Tononi, G. (2021). Of maps and grids. *Neurosci Conscious, 2021*(2), niab022. doi:10.1093/nc/niab022

Graziano, M. S. A. (2022). A conceptual framework for consciousness. *Proceedings of the National Academy of Sciences, 119*(18), e2116933119. doi:doi:10.1073/pnas.2116933119

Graziano, M. S. A., & Webb, T. W. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology, 6*(500). doi:10.3389/fpsyg.2015.00500

Hacking, I. (1982). Experimentation and Scientific Realism. *Philosophical Topics, 13*(1), 71-87. Retrieved from http://www.jstor.org/stable/43153910

Hacking, I. (1988). Philosophers of Experiment. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1988*, 147-156. Retrieved from http://www.jstor.org/stable/192879

Haun, A. M., Tononi, G., Koch, C., & Tsuchiya, N. (2017). Are we underestimating the richness of visual experience? *Neuroscience of Consciousness, 2017*(1). doi:10.1093/nc/niw023

Henderson, L., Goodman, N., Tenenbaum, J., & Woodward, J. (2010). The Structure and Dynamics of Scientific Theories: A Hierarchical Bayesian Perspective*. *Philosophy of Science, 77*(2), 172-200. doi:10.1086/651319

Hesse, M. (1974). *The Structure of Scientific Inference*. Berkeley: University of California Press.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. Chicago, IL, US: Open Court Publishing Co.

Lakatos, I. (1968a). Changes in the Problem of Inductive Logic. In I. Lakatos (Ed.), *Studies in Logic and the Foundations of Mathematics* (Vol. 51, pp. 315-417): Elsevier.

Lakatos, I. (1968b). Criticism and the Methodology of Scientific Research Programmes. *Proceedings of the Aristotelian Society, 69*, 149-186.

Lakatos, I. (1974). The role of crucial experiments in science. *Studies in History and Philosophy of Science Part A, 4*(4), 309-325. doi:https://doi.org/10.1016/0039-3681(74)90007-7

Lakatos, I. (1976). Falsification and the Methodology of Scientific Research Programmes. In S. G. Harding (Ed.), *Can Theories be Refuted? Essays on the Duhem-Quine Thesis* (pp. 205-259). Dordrecht: Springer Netherlands.

Lakatos, I., & Musgrave, A. (1970). *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965* (I. Lakatos & A. Musgrave Eds. Vol. 4). Cambridge: Cambridge University Press.

Lakatos, I., & Zahar, E. (1975). Why Did Copernicus' research program supersede Ptolemy's? In S. W. Robert (Ed.), *The Copernican Achievement* (pp. 354-383). Berkeley: University of California Press.

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494-501. doi:10.1016/j.tics.2006.09.001

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience, 1*(3), 204-220. doi:10.1080/17588921003731586

Lau, H. (2022). *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*: Oxford University Press.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences, 15*(8), 365-373. doi:10.1016/j.tics.2011.05.009

Lenharo, M. (2023). Decades-long bet on consciousness ends — and it's philosopher 1, neuroscientist 0. *Nature, 619*. doi:https://doi.org/10.1038/d41586-023-02120-8

Lipton, P. (1990). Prediction and prejudice. *International Studies in the Philosophy of Science, 4*(1), 51-65. doi:10.1080/02698599008573345

Maher, P. (1988). Prediction, Accommodation, and the Logic of Discovery. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 1988*, 273 - 285.

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron, 105*(5), 776-798. doi:10.1016/j.neuron.2020.01.026

Massimini, M., Boly, M., Casali, A., Rosanova, M., & Tononi, G. (2009). A perturbational approach for evaluating the brain's capacity for consciousness. In S. Laureys, N. D. Schiff, & A. M. Owen (Eds.), *Progress in Brain Research* (Vol. 177, pp. 201-214): Elsevier.

Massimini, M., & Tononi, G. (2018). *Sizing up Consciousness. Towards an objective measure of the capacity for experience* (F. Andersen, Trans.). Oxford: Oxford University Press.

Mayo, D. G. (1991). Novel Evidence and Severe Tests. *Philosophy of Science, 58*(4), 523-552. Retrieved from http://www.jstor.org/stable/188479

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*: University of Chicago Press.

McKenzie, C. R. M., & Amin, M. B. (2002). When wrong predictions provide more support than right ones. *Psychonomic Bulletin & Review, 9*(4), 821-828. doi:10.3758/BF03196341

Melloni, L. (2022). On keeping our adversaries close, preventing collateral damage, and changing our minds. Comment on Clark et al. *Journal of Applied Research in Memory and Cognition, 11*(1), 45-49. doi:10.1037/mac0000009

Melloni, L., Mudrik, L., Pitts, M., Bendtz, K., Ferrante, O., Gorska, U., . . . Tononi, G. (2023). An adversarial collaboration protocol for testing contrasting predictions of global neuronal workspace and integrated information theory. *PLOS ONE, 18*(2), e0268577. doi:10.1371/journal.pone.0268577

Melloni, L., Mudrik, L., Pitts, M., & Koch, C. (2021). Making the hard problem of consciousness easier. *Science, 372*(6545), 911-912. doi:doi:10.1126/science.abj3259

Musgrave, A. (2023). Imre Lakatos. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Spring 2023 ed.): Metaphysics Research Lab.

Musgrave, A., & Mayo, D. G. (2009). Revisiting Critical Rationalism. In A. Spanos & D. G. Mayo (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (pp. 88-124). Cambridge: Cambridge University Press.

Nozick, R. (1983). Simplicity as fall-out. In L. Cauman, I. Levi, C. D. Parsons, & R. Schwartz (Eds.), *How Many Questions?* : Hackett.

Nunan, R. (1984). Novel facts, Bayesian rationality, and the history of continental drift. *Studies in History and Philosophy of Science Part A, 15*(4), 267-307. doi:https://doi.org/10.1016/0039-3681(84)90013-X

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol, 10*(5), e1003588. doi:10.1371/journal.pcbi.1003588

Parr, T., Corcoran, A. W., Friston, K. J., & Hohwy, J. (2019). Perceptual awareness and active inference. *Neuroscience of Consciousness, 2019*(1). doi:10.1093/nc/niz012

Phillips, I. (2011). Perception and Iconic Memory: What Sperling Doesn't Show. *Mind & Language, 26*(4), 381-411. doi:https://doi.org/10.1111/j.1468-0017.2011.01422.x

Popper, K. R. (1959). *The logic of scientific discovery*. Oxford, England: Basic Books.

Popper, K. R. (1968). *Conjectures and Refutations: The Growth of Scientific Knowledge*: Harper & Row.

Quine, W. V. O. (1951). Two Dogmas of Empiricism. *Philosophical Review, 60*(1), 20–43.

Sarasso, S., Casali, A. G., Casarotto, S., Rosanova, M., Sinigaglia, C., & Massimini, M. (2021). Consciousness and complexity: a consilience of evidence. *Neuroscience of Consciousness*. doi:10.1093/nc/niab023

Scerri, E. R., & Worrall, J. (2001). Prediction and the periodic table. *Studies in History and Philosophy of Science Part A, 32*(3), 407-452.

Seth, A. K. (2023). Finding the Neural Correlates of Consciousness Is Still a Good Bet. Retrieved from https://nautil.us/finding-the-neural-correlates-to-consciousness-is-still-a-good-bet-352054/

Seth, A. K., & Bayne, T. (2022). Theories of consciousness. *Nature Reviews Neuroscience*. doi:10.1038/s41583-022-00587-4

Signorelli, C. M., Szczotka, J., & Prentner, R. (2021). Explanatory profiles of models of consciousness - towards a systematic classification. *Neuroscience of Consciousness, 2021*(2). doi:10.1093/nc/niab021

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1-29. doi:10.1037/h0093759

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neurosci, 5*, 42. doi:10.1186/1471-2202-5-42

Tononi, G. (2012). Integrated information theory of consciousness: an updated account. *Arch Ital Biol, 150*(4), 293-329. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23802335

Tononi, G. (2015). Integrated information theory. *Scholarpedia*. doi:doi::10.4249/scholarpedia.4164

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci, 17*(7), 450-461. doi:10.1038/nrn.2016.44

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philos Trans R Soc Lond B Biol Sci, 370*(1668). doi:10.1098/rstb.2014.0167

Vilas, M. G., Auksztulewicz, R., & Melloni, L. (2022). Active Inference as a Computational Framework for Consciousness. *Review of Philosophy and Psychology, 13*(4), 859-878. doi:10.1007/s13164-021-00579-w

Wiese, W., & Friston, K. J. (2021). The neural correlates of consciousness under the free energy principle: From computational correlates to computational explanation. *Philosophy and the Mind Sciences, 2*. doi:10.33735/phimisci.2021.81

Worrall, J. (1985). Scientific Discovery and Theory-Confirmation. In J. C. Pitt (Ed.), *Change and Progress in Modern Science: Papers Related to and Arising from the Fourth International Conference on History and Philosophy of Science* (pp. 301-331). Dordrecht: Reidel.

Worrall, J. (1989). Fresnel, Poisson and the white spot: The role of successful predictions in the acceptance of scientific theories. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The Uses of*

Experiment: Studies in the Natural Sciences (pp. 135-157). Cambridge: Cambridge University Press.

Worrall, J., & Mayo, D. G. (2009). Theory Confirmation and Novel Evidence. In A. Spanos & D. G. Mayo (Eds.), *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science* (pp. 125-169). Cambridge: Cambridge University Press.

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

Zahar, E. (1973). Why Did Einstein's Programme Supersede Lorentz's? (I). *The British Journal for the Philosophy of Science*, *24*(2), 95-123. Retrieved from http://www.jstor.org/stable/686604