This is a peer-reviewed manuscript submitted to Melloni, L. & Olcese, U. (Eds.) (Forthcoming), The Scientific Study of Consciousness– Experimental and Theoretical Approaches. Springer Nature.

The Philosophy of Science of Consciousness Science

Niccolò Negro[1]
niccolonegro@tauex.tau.ac.il
https://orcid.org/0000-0002-1561-799X

Andrew W. Corcoran[2]
https://orcid.org/0000-0002-0449-4883

Liad Mudrik[1,3,4]
https://orcid.org/0000-0003-3564-6445

Jakob Hohwy[2]
https://orcid.org/0000-0003-3906-3060

[1]School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel.

[2]Monash Centre for Consciousness and Contemplative Studies, Monash University, Melbourne, Australia.

[3]Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel.

[4]Canadian Institute for Advanced Research (CIFAR), Brain, Mind, and Consciousness Program, Toronto, ON, Canada.

This chapter surveys some potential contributions of philosophy of science to the scientific study of consciousness. Given the unique challenges consciousness poses as both a subjective and objective phenomenon, philosophy of science can offer conceptual tools for clarifying definitions, establishing methodological frameworks, and guiding theory comparison and assessment. By integrating philosophical perspectives on general philosophy of science with specific debates within the science of consciousness, this chapter aims to demonstrate how philosophy of science can support consciousness researchers in navigating the complexities of their field and accelerating its progress. We suggest that a promising route to make progress in consciousness science is to combine three complementary and mutually reinforcing strategies: the empirical strategy, the confirmational strategy, and the metatheoretical strategy.

Consciousness; Philosophy of Science; Scientific Progress; Adversarial Collaboration; Confirmation Theory.

Introduction: Leveraging philosophy of science for progress in consciousness science

The relationship between modern science and philosophy of science is a fascinating and hotly debated topic. While some might adhere to the view (often attributed to Richard Feynman) that philosophy of science is as useful to scientists as ornithology is to birds, others might instead side with Einstein's idea that philosophical reflection is "the mark of distinction between a mere artisan or specialist and a real seeker after truth" (Einstein, 1944; cited in Laplane et al., 2019).

The scope of this chapter is not to resolve this perennial debate, but, more modestly, to flag that philosophy of science might indeed be quite useful to the specific field of consciousness science (for similar views, see Buccella, Maoz, & Mudrik, 2024; Kirkeby-Hinrup, 2024a). This by no means entails that philosophy of science should guide the practice of consciousness science, but only that philosophical insights and analyses might help situate, finesse and justify scientists' various stances on foundational issues in consciousness science. We believe that this is the case both because of the particular epistemic situation consciousness scientists find themselves in, given the nature of the phenomenon they are studying, and because of the state of the field.

First, consciousness science is distinctive: even for naturalist scholars inclined to think that consciousness science should proceed according to the standard methods of other natural sciences (like physics, chemistry, biology, etc.), consciousness poses an epistemic problem (i.e., with respect to how it should be studied and explored), since it figures both as object of study from the third-person perspective and as an indispensable subjective component from the first-person perspective.

This unusual relationship between subjective and objective, or between the first-person perspective and the third-person perspective, has been emphasized by different philosophical traditions (Chalmers, 2004; T. Nagel, 1974; Searle, 2017; Varela, 1996), translating into scientific discussions on whether this unique epistemic status of consciousness as a scientific target requires revolutionary methodologies (Block, 2007; Cohen & Dennett, 2011; Doerig, Schurger, Hess, & Herzog, 2019; Lamme, 2006; Tononi & Koch, 2015). In this context, philosophy of science can help provide the conceptual tools for defining what a scientific explanation is and for determining the limits of its applicability. Moreover, a deep look at the philosophy of science can reveal patterns of explanation and methodological strategies that can enrich the repertoire of consciousness scientists.

Second, consciousness science is a relatively young scientific field. Although it stems from the convergence of more established fields (like cognitive neuroscience, psychology, neurobiology, etc.), each with quite clearly defined problems, methods and heuristics, consciousness science as an autonomous and independent discipline has not inherited a well-established methodological and conceptual foundation (for reviews of its historical roots, see Baars, 2009; LeDoux, Michel, & Lau, 2020; Michel, 2019). In this context, different schools of thought interpret the explanandum (i.e., consciousness; the phenomenon that needs to be explained) in different ways, and build their explanans (i.e., what does the explaining) upon different explanatory constructs (Fazekas, Cleeremans, & Overgaard, 2024; Signorelli, Szczotka, & Prentner, 2021) and different methodological approaches (Yaron, Melloni, Pitts, & Mudrik, 2022). Philosophy of science can help by better defining the explanandum, and by providing rigorous conceptual frameworks for connecting the explanandum to a specific explanans. Moreover, a perspective that integrates both the history and the philosophy of science can inform consciousness scientists on the development of scientific disciplines that have previously gone through a similar trajectory, so that consciousness science can learn from these disciplines and accelerate its progress.

Third, because of this state of the field, many theories and frameworks ‘compete’ for becoming the main paradigm in consciousness science. Philosophy of science can help approach the problem of theory comparison and theory assessment, and can accordingly be a useful conceptual companion for empirical work done to test and compare theories of consciousness.

In summary, philosophy (and the history) of science can aid consciousness scientists in navigating the rapid expansion of their relatively young field by providing conceptual tools to: I) clearly define the explanandum and the explanatory strategies to address the unique epistemic status of consciousness; II) investigate the relationship between theories and models to better understand the limits of scientific theories; III) assess the scientific status of theories of consciousness; and IV) compare and evaluate theories of consciousness. Below we focus on how this could be done for each one of these domains. Different approaches in philosophy of science might provide different answers to the problems we raise; our goal here is not to claim that these are the necessary solutions, but to showcase the type of contributions philosophy of science can offer to the scientific study of consciousness.

1. Explanandum and Explanans

There has been ongoing discussion about the definition of consciousness, both at the conceptual and at the operational level (Crick & Koch, 1990, 1998; Searle, 1998, 2000; Zeman, 2001). Here, philosophy of science plays a major role. To demonstrate, we will focus on one of the main points of disagreement in consciousness science: the distinction between access and phenomenal consciousness (Block, 1995), and whether it is a scientifically legitimate distinction or not. In a nutshell, access consciousness refers to the aspects of consciousness that are available to cognitive processes, and can be used for reasoning, decision-making, action planning, and so on. On the other hand, phenomenal consciousness refers to the qualitative feeling that accompanies our mental life. Note that by giving this example, we are not embracing this distinction, nor do we argue that it should guide the scientific study of consciousness; our goal is merely to use this issue to illustrate that whatever stand one takes with respect to this distinction, philosophy of science is pivotal for understanding the impact each stand has on the epistemological status of the scientific investigation.

Some argue that, from a scientific point of view, this distinction is meaningless, because phenomenal consciousness is not scientifically tractable (Cohen & Dennett, 2011; Dennett, 1995; Naccache, 2018; Naccache & Dehaene, 2007). In fact, the thought goes, scientific data are by definition third-person data, and therefore, if we want to study consciousness in a scientifically respectable way, any first-person experience must be translated into a third-person description. For some, this amounts to denying the existence of phenomenal consciousness as a scientific phenomenon at all (Dennett, 2016; Frankish, 2016; Irvine, 2012, 2017), maintaining instead that access consciousness is the only real, and scientifically meaningful, phenomenon of interest. For others, however, the requirement of translating consciousness into third-person data just amounts to studying phenomenal consciousness through access consciousness (Herzog, Schurger, & Doerig, 2022; Naccache, 2018). In this latter sense, the distinction between access and phenomenal consciousness might echo the distinction in the philosophy of science literature between data and phenomena (Bogen & Woodward, 1988), where phenomena are thought to be regularities in nature extracted and inferred from experimental observations (i.e., data). In the natural sciences, it is not uncommon to infer a phenomenon (or an entity) from another phenomenon, more directly related to the observed dataset. For example, the existence of unobservable negatively charged particles, now called electrons, was inferred from the phenomenon of deflection of cathode rays, which, in turn, was directly inferred from data such as the locations in which a beam of cathode rays intercepted a plate at the end of the tube. So, scholars who maintain that phenomenal consciousness can be studied only through access consciousness and third-person data, but are reluctant to embrace an eliminativist position, still conceive phenomenal consciousness as the ultimate target of consciousness science. It just needs to be inferred from third-person data afforded by more directly available phenomena. This methodological approach to consciousness does not require any revolutionary strategy, and claims that consciousness science, done in this way, is perfectly continuous with the standard methodology of the natural sciences.

However, an opposite view states that the distinction between data and phenomena collapses because of the unique epistemic status of consciousness: what is given immediately to the subject is both the datum and the phenomenon that needs to be explained, and therefore consciousness science must start from first-person data to make sure that it is targeting the right phenomenon. According to this view, eliminating phenomenal consciousness just means eliminating the phenomenon we must explain, while studying phenomenal consciousness through access consciousness risks including confounding factors that might hinder our understanding of consciousness as such.

The question for this approach, then, is how to proceed in a scientifically legitimate way once we have established the necessity of these scientifically unorthodox first-person data. One possible approach of this sort is the integrated information theory (IIT) (Albantakis et al., 2023; Ellia et al., 2021). IIT maintains that a scientific explanation of consciousness can be derived from consciousness itself by positing operational postulates designed to explain how the physical world must be to underpin the essential properties of consciousness (see Bayne, 2018; Merker, Williford, & Rudrauf, 2021; Negro, 2022a, 2022b; Signorelli, Cea, & Prentner, 2024 for discussions of the scientific legitimacy of this “bootstrapping” process).

A “middle way” view between the traditional third-person perspective approach and the revolutionary first-person perspective approach has been suggested by Seth (2021) and Hohwy and Seth (2020). According to this approach, it is reasonable to start with a general framework for brain functioning, which is not necessarily a theory of consciousness (but more of a framework *for* consciousness science), and ask how this framework can explain and predict features of consciousness. In this sense, both the third and the first person perspectives are taken seriously, and figure in the explanatory strategy for explaining consciousness: the first-person perspective fixes, via phenomenological investigation, the properties of consciousness that need to be explained, while the third-person perspective introduces a framework for brain functioning that can account for them. However, it is still unclear whether a full-fledged theory of consciousness can be derived from this sort of approach (for an example of how this could work, see Whyte et al., in preparation).

Here again, philosophy of science can help the scientific investigation of consciousness: by elucidating the different epistemological commitments of the different approaches, it can situate the discussion better within a wider context, providing scientists with additional tools to evaluate the potential costs and benefits of the various approaches in the scientific study of consciousness.

2. Theories and models of consciousness

When assessing the limits of scientific theories, it is important to clarify what constitutes a theory. In this regard, one of the most discussed and relevant distinctions in philosophy of science is between theoretical and observational statements. Observational statements are often defined as featuring terms that refer to empirically observable objects (e.g., "subjects' response time was faster in condition 1 than in condition 2"), while theoretical statements as featuring theoretical constructs that cannot be directly observed (e.g., "consciousness requires information to be maintained in the global workspace"). However, this is a controversial distinction in philosophy of science, since it is debatable whether a clear line between theory and observation can be drawn. For example, observations of a distant planet with the help of a telescope count as reliable data only if one relies on optical theories that explain how the telescope works. Hanson (1965) captured this fact by arguing that observation is always "theory-laden". Thus, one of the central tasks for philosophers of science is to explain how theoretical terms get their meaning, and to elucidate the exact relationship between theoretical and observational statements (Carnap, 1956; Nagel, 1961; Quine, 1951 – see Vorms, 2018 for a discussion). This helps provide an account of what constitutes a theory.

This is especially important for consciousness science because i) a multitude of theories have been proposed, varying in scope and breadth (Francken et al., 2022; Seth & Bayne, 2022;

Yaron et al., 2022); and ii) fragmentation of the field has motivated, as one of the possible ways to accelerate progress, adversarial collaborations designed to empirically test theories of consciousness against each other (e.g., Cogitate et al., 2023). Given the focus on theory testing, adversarial collaboration seems to require a clear way to tell how empirical tests, targeting observational statements, reverberate on theoretical statements. This further requires taking a stance on the question of how theories are structured, and therefore on the relationship between theories and models, because this relationship is ultimately what constitutes a theory. However, it is doubtful whether this clarity is currently achieved in consciousness science. This is where philosophy of science can again advance the field and help assess the contribution of projects like adversarial collaborations to our understanding of consciousness.

A popular view in philosophy of science is that theories are collections of models, where models are defined as structures that stand between the theory and the world, satisfying some general theoretical principles to represent a phenomenon and to allow for predictions and explanations (Craver, 2002). It is standard scientific practice to focus more on models than theories, since models provide an idealized and simplified setting to intervene on a phenomenon and therefore to understand it better (Cartwright, 1999; Godfrey-Smith, 2006). Theories would be, in this case, general constructions that abstract away from specific modelling features, and provide a more general, non-specific understanding of a phenomenon (see Vorms, 2018 and Craver, 2002 for discussions).

Importantly, models can be logical models or representational models. Logical models are models *of a theory*, translating its claims to testable predictions such that they constitute an interpretational tool of a theory. Representational models, instead, are models *of an actual natural phenomenon*, such that they are aimed at representing the phenomenon itself, and not the theory describing it (for analyses of types of models, see Suppes, 1969). Typically, a single model has both logical and representational aspects (Hesse, 1966).

This distinction is helpful for assessing tests of theories of consciousness: it seems that specific models (with both their logical and representational aspects) can facilitate the formulation of *specific* predictions of a theory, which are then empirically tested. But if a theoretical prediction is formulated through a model, and filtered by it, then the prediction might depend on properties of the model that are not directly driven from the theory that motivates the model. As Frigg (2022) puts it: “We can infer from the billiard ball model [of an ideal gas] that when a gas molecule collides with the wall of the vessel the angle of incidence is equal to the angle or reflection, but we cannot infer that molecules are coloured and have numbers written on them" (Frigg, 2022, p. 52).

The problem in consciousness science is that the models adopted to explain and predict consciousness-related phenomena might be too sensitive to their own specific properties, rendering the specific predictions derived from these models only partially informative of the theory that generates them. Thus, despite models in consciousness science being *representational* models, their *logical* nature (i.e., how they connect to the core principles of the theory) might be unclear.

To exemplify this point, take the computational model Dehaene et al. (2003) developed for modelling the link between an attentional blink experiment and neurophysiological data. The model shows how entrance into the global workspace is signalled by an all-or-none dynamics of brain activity (called "ignition"), demonstrating that this type of activity correlates with conscious perception of a stimulus. Vice versa, blinked stimuli (i.e., stimuli that are presented immediately after a previous stimulus) are not consciously perceived and only generate bottom-up activity which does not reverberate globally, but fades away quickly. This model is taken to support the global workspace theory of consciousness (GNWT) (Dehaene & Naccache, 2001; Mashour, Roelfsema, Changeux, & Dehaene, 2020) insofar as it shows that information must be globally maintained in a workspace in order to be consciously perceived, and that if the workspace is occupied by one item (i.e., the first stimulus in the attentional blink paradigm), it will be unlikely that a second item (i.e., the second stimulus in the attentional blink paradigm) is consciously perceived.

The model simulates neuronal dynamics, membrane potentials, sodium and potassium currents, and so on (Dehaene et al., 2003, p. 8521). However, these aspects of the model are not driven by the theory, and the theory's success does not depend on them: it is possible to build a slightly different model with different parameters and different parameter values, which would still be GNWT-inspired, GNWT-supporting or at least GNWT-compatible. The key issue is that, under different computational models, the type of neural dynamics taken to be the neural signature of ignition might be slightly different, and might not be indexed by the P300 component, as originally hypothesized by GNWT proponents (see Koivisto, Salminen-Vaparanta, Grassini, & Revonsuo, 2016; Mashour et al., 2020; Pitts, Metzler, & Hillyard, 2014 for discussions). This is important, because many empirical predictions used to test GNWT employ the theoretical construct of ignition and its mechanistic implementation (Cogitate et al., 2023). But if the properties of this construct vary depending on how we build a specific model, then the empirical predictions related to this construct will also be model-specific predictions, and might not directly speak to the higher-level theory we want to test. This is of course not unique to GNWT: similar examples can be given for other theories. The overall point is that without a clear view of how models are logically related to theories of consciousness, the empirical tests derived from those models are undermined as tests for *theories* of consciousness. This boils down to taking a stance towards the question of what constitutes a theory, and how it relates to models, an often forgotten and underappreciated issue that both scientists and philosophers working in consciousness science should address. This, again, is where philosophy of science can be highly instrumental to the study of consciousness: both in examining existing models and their relations to specific theories, and in assessing the relations between theories, models, and empirical tests.

3. The scientific status of theories of consciousness

When evaluating theories of consciousness, a key issue is if the theory should be considered a *scientific* theory or not – which, again, is a question of philosophy of science. This is particularly important in consciousness science, since the problem of consciousness has for centuries been a mainly philosophical problem, and it is typical of nascent scientific fields to be motivated more by theoretical speculations than empirical evidence (Kuhn, 1970).

In philosophy of science, the problem of discriminating what counts as science and what does not is called the “demarcation problem” (see Laudan, 1983 for a critical analysis). This problem features in several specific debates in the science of consciousness. For example, Doerig et al. (2019) claimed that IIT and the recurrent processing theory (Lamme, 2010) are either false or unscientific because they are unfalsifiable (for replies, see Kleiner, 2020; Tsuchiya, Andrillon & Haun, 2020; Negro, 2020; Usher, 2021; for general discussions, see Kleiner & Hoel, 2021; Hanson & Walker, 2021; Herzog et al., 2022; Usher, Negro, Jacobson & Tsuchiya, 2023). This argument is based on Popper’s influential analysis of the demarcation problem according to which a theory is scientific if and only if it is falsifiable (Popper, 1959).

Without taking a stand on the argument itself here, we note that more recent discussions in philosophy of science questioned whether this popular Popperian view is indeed the best way to carve the distinction between science and non-science, since it seems too strict and ends up labelling legitimate scientific endeavours pseudoscientific (string theory and the Many-World interpretation of quantum mechanics, for example; see Hansson, 2006 for a general discussion and analysis).

The discussions within the field of consciousness science can accordingly benefit from these more recent views on the demarcation problem. Here, we present three different approaches to this problem and explain how they could reflect on the scientific study of consciousness.

One such approach is based on the view of Imre Lakatos. Recently, Negro (2020, 2024) suggested that this view is more appropriate in understanding the scientific status of theories of consciousness than the Popperian one. This sophisticated version of falsificationism focuses on the progressivity of research programmes and allows for theoretical revisions based on incoming evidence. The basic Lakatosian tenet is that a theory is composed of core ideas and various belts of auxiliary hypotheses (i.e., hypotheses that are assumed to be true for the theory to hold). A theory is never tested directly at its core, but always through some combination of auxiliaries. So, if a prediction is falsified, the theorist can simply modify some of the auxiliaries while maintaining the truth of the core of the theory. This process generates a research programme. According to Lakatos, research programmes, not theories, are the right units of analysis when discussing the demarcation problem. A research programme is progressive if modifications and revisions lead to more empirically testable predictions, while it degenerates if they do not (Lakatos, 1976).

The Lakatosian strategy, however, does not clarify when exactly it is rational to abandon one research programme in favour of another. Because of this, it seems possible that non-falsificationist accounts could be advanced to deal with this problem.

Another approach denies the problem itself, stating that that there is no such thing as “the scientific method”, and therefore there is no clear way to distinguish between science and non-science (Feyerabend, 1988). If this “methodological anarchism” is correct, then the very question of whether some theories of consciousness deserve the title of “scientific” theories while others do not, becomes meaningless. Science, according to this view, is mostly driven by personal, social, and subjective factors, not by objective criteria.

Independently of whether one endorses such a radical view, one of the merits of methodological anarchism consists in emphasizing the social situatedness of scientists, suggesting that social aspects of scientific practice might play a major role in the development of science. In consciousness science specifically, this has also been flagged by scholars who have included a similar perspective in a substantially less radical framework, inspired by Kuhn (1970). In fact, both Merker and colleagues (2021) and Evers and colleagues (2024) have noticed that consciousness science displays many of the properties that Kuhn thought were characteristic of the "prehistory" of a scientific discipline, like the competition of different schools to impose themselves on others while disagreeing on methodologies, heuristics, and even metaphysical assumptions. According to Kuhn, normal science is dominated by a paradigm, namely a set of beliefs, methodologies, and techniques that constitute the "disciplinary matrix" within which scientists accumulate knowledge via puzzle solving. When the problems within this paradigm become untenable, normal science enters a revolutionary phase. In this case, a novel paradigm is proposed to solve the problems that the old paradigm was struggling with, and when the scientific community shifts from the disciplinary matrix provided by the old paradigm to the new matrix provided by the new paradigm, a scientific revolution has been accomplished.

In the case of consciousness science, scholars who have adopted this Kuhnian lens (Evers et al., 2024; Merker et al., 2021) have concluded that the discipline is not mature enough to be considered as a normal science. An analogy might be drawn with how Kuhn presented the pre-scientific (in his terms, "pre-paradigmatic") state of physical optics before Newton's work. In this context, many groups and schools of thought were competing to establish the right theory of light, but they tried to do so through various different methods, adopting different definitions of the target phenomenon, and relying on different metaphysical assumptions: "Each of the corresponding schools derived strength from its relation to some particular metaphysic, and each emphasized, as paradigmatic observations, the particular cluster of optical phenomena that its own theory could do most to explain. Other observations were dealt with by *ad hoc* elaborations, or they remained as outstanding problems for further research" (Kuhn, 1970, p. 13).

This seems indeed reminiscent of the current state of consciousness science, given that, as Yaron et al. (2022) show, prominent theories of consciousness are mostly developed independently and seek validation by experimental methodologies that are well-suited to verify the predictions a particular theory wants to verify. One of the crucial findings by Yaron et al. is that whether an experiment supports a given theory can be predicted just by looking at the experimental methodology adopted, even without considering the results (fig. 4 in Yaron et al., 2022). Like pre-paradigmatic theories of light, consciousness theories seem to emphasize as paradigmatic only those types of phenomena that each theory is well-suited to explain.

In this context, the question of whether a theory of consciousness is scientific or not might be simply premature, because the discipline itself is still at a pre-scientific stage.

These considerations may help the debate shift from the issue of the scientific status of theories of consciousness to the issue of how to make and track scientific progress in the neuroscience of consciousness. We turn to this issue next.

4. Making progress in consciousness science

In the previous sections, we demonstrated where philosophy of science can illuminate key issues that are important for understanding claims made in the field of consciousness: with respect to the explanandum, to the status of the theories, and to the ability to test them. We further suggested that the field might still be at a pre-paradigmatic stage, in Kuhnian terms. If so, how can we move forward and make genuine scientific progress? Although this question might not even make sense within a Kuhnian framework, since for Kuhn science is characterised by paradigm shifts without progress, there are alternative philosophical accounts that can help us understand what is meant by scientific progress, and whether it is a matter of accumulation of knowledge, closeness to truth, or increased problem-solving effectiveness (Bird, 2007; Howson & Urbach, 1989; Laudan, 1977; Popper, 1968; Shan, 2019). Here too, philosophy of science might be instrumental to consciousness science, as we show below.

Conceptually, several strategies for advancing consciousness science have been proposed, which we regard as mainly complementary, rather than mutually exclusive. We categorize them into three classes: empirical, confirmational, and metatheoretical, and show how they can be implemented in consciousness science. Each has its own advantages and disadvantages, as we explain below, but they also hold the potential for making true progress in the field of consciousness science, especially when they are taken as complementary such that they can inform each other.

Empirical attempts see scientific progress to be driven mainly by accumulation of empirical evidence. This strategy prescribes that consciousness science will eventually move to the paradigmatic stage by testing empirical predictions of scientific theories: a paradigm will spontaneously emerge based on which theory does better in the empirical arena and only at that point will consciousness science be a normal science, in the Kuhnian sense. *Confirmational approaches* seek to pair empirical work with a confirmation theory account that measures to what extent a certain experimental result confirms or disconfirms a theory. Finally, *metatheoretical attempts* try to compare the central constructs and measures of various theories of consciousness to determine their comparability and potential avenues for integration.

The empirical strategy

The most prominent example of the empirical strategy is provided by the series of adversarial collaborations between competing theories of consciousness (Cogitate et al., 2023; Melloni, 2022; Melloni et al., 2023; Melloni, Mudrik, Pitts, & Koch, 2021). In these collaborations, proponents of different theories co-design experiments aimed at testing contrasting predictions of their theories, in concert with independent and theory-neutral experimentalists who run the experiments (for general discussions, see Clark, Costello, Mitchell, & Tetlock, 2022; Clark & Tetlock, 2023; Cowan et al., 2020; Kahneman, 2003). The driving force of this strategy is to put theories under severe tests that could challenge and put pressure on specific predictions, contrary to the standard practice of consciousness science (see Melloni, 2022), of verifying predictions through methods that are well-suited to confirm those predictions.

However, it is debatable whether every experiment born in this adversarial context counts as a severe test. One of the most refined formulations of severe testing comes from Deborah Mayo's work (Mayo, 1991). Mayo's notion of a severe test is defined as a test designed to confirm a hypothesis (i.e., demonstrate concordance between the hypothesis and the data) that would have most probably not passed the test, were the hypothesis false. That is, for a test to be severe, it should be very probable for a true hypothesis to pass it, and very improbable for a false hypothesis to pass it. Therefore, if a hypothesis fails the test, we have good reasons to think the hypothesis is false. In order to design severe tests, it is thus crucial to investigate and eliminate the various reasons why the hypothesis might be wrong.

For example, imagine we design an experiment to test the hypothesis that a new drug helps with anxiety. We give this new drug to a dozen undergraduate students who self-report being anxious during the exam period, and ask them to track their mood for a month. Even if results showed a significant reduction in stress levels, the hypothesis is not confirmed, because the test was not severe: the sample size is too small, there is no control group, self-reportability might be an inaccurate measurement, and there are many other factors that might equally explain the result (e.g., perhaps students just felt less anxious because the exam period ended, perhaps some started practicing meditation, or the results might just reflect a placebo effect, etc.). In a nutshell, it is quite probable for the hypothesis to pass the test even if it was false, and therefore, according to Mayo's account, the test does not count as a severe test, and we cannot say that we gained much, if any, scientific knowledge from it.

Following this definition of severe testing, Negro (2024) has argued that some (but not all) of the predictions in the first experiment of the Cogitate adversarial collaboration between IIT and GNWT are not severely tested, since the probability of the predictions passing the test is quite high even if IIT and GNWT were false. Yet even under this scenario, we hold that the adversarial collaboration between IIT and GNWT is informative and contributes to genuine scientific progress in the science of consciousness: while the confirmed hypotheses might not be enough to support the theories, the failed predictions are highly meaningful. For example, in the first Cogitate experiment, stimuli were designed to evoke conscious experiences accompanied by strong neural responses, such that null results cannot be explained away by appealing to the weakness of the evoked signals. Moreover, three different neuroscientific techniques that compensate for each other's limitations were used, together with an optimization phase where the theorists are allowed to try different analytic strategies (for a comprehensive discussion see Cogitate et al., 2023). Given this setup, failed predictions still put considerable pressure on the theories they derive from, and this is by itself relevant information that can contribute to progress in consciousness science.

This example suggests that severe testing might be too aspirational as the only way to evaluate hypotheses, and the claim that scientific knowledge can be accumulated only through severe testing might be too strict. The question, then, is how to make sense of the idea that scientific progress can be achieved even through tests that are not severe. The Cogitate approach, of focusing on failed predictions as opposed to confirmed ones, is one approach. Another approach is the confirmation-theoretic one, which scores the degree of confirmation an empirical result bears with respect to a theory.

The confirmational strategy

We showed that empirical testing should best be complemented by a clear account of how a certain result challenges or disconfirms a theory. According to the confirmational strategy, progress in consciousness science can be made by developing such a confirmation-theoretic account. In this regard, a promising confirmation theoretic framework for adversarial collaborations in consciousness science has been advanced by Corcoran and collaborators (2023; see also, for another Bayesian approach, Kirkeby-Hinrup, 2024).

According to this view, adversarial collaborations should be conceived and interpreted through the lens of Bayesian belief updating; that is, credence about a certain theory should be modified according to Bayes' conditionalization rule $P(A \mid B) = (P(A \mid B)P(A))/P(B)$. This involves asking adversaries to agree on one or more experiments capable of generating data that can discriminate between their theories, and to commit to specific predictions ('Bayesian bets') about the outcome of these experiments on critical parameters of interest. Then, statistical (generative) models are constructed that fix the relevant parameters pertaining to each experiment. Crucially, each model is equipped with different prior distributions on key parameters of interest, where such priors specify plausible parameter values under each theorist's Bayesian bet about the outcome of a given experiment. Finally, each model is fitted to the empirical data garnered during the experiment to obtain an estimate of model evidence (i.e., marginal likelihood). This quantity scores how well the fitted model captures the data (model accuracy) versus the degree of belief updating required to fit the model to the data (model complexity). In this sense, estimating model evidence allows us to consider how each model fares in terms of the interplay between predictive accuracy and parsimony (Forster & Sober, 1994), two criteria that are often discussed as crucial theoretical virtues that should guide theory-choice (Keas, 2018; McMullin, 2013).

Once models have been fitted to empirical observations, Bayesian model comparison can be performed to compare the evidence accrued under each model. If one theorist's initial Bayesian bet about the outcome of the experiment was accurate, the model encoding this prediction will accrue more evidence than alternative models equipped with less-accurate priors, assuming equal degrees of model complexity. Conversely, a simpler model will accrue more evidence in its favour than more-complex alternatives that make equally-accurate predictions. This means that theorists who stake more precise Bayesian bets (i.e., commit to prior predictions that assign higher probabilities to a small set of parameter values) stand to ‘win’ more evidence than those who prefer to ‘hedge’ their bets (i.e., commit to priors that distribute probabilities more evenly over a broad range of values) – but also stand to ‘lose’ more evidence (relative to their competitor) if their predictions turn out to be wrong. In this way, it is possible to score how well a theory performs according to the accuracy and complexity of its associated Bayesian bets, and thus to rank competing theories by the amount of evidence their models accumulate:

> It should be noted that evidence is only meaningful in a relative sense. In other words, one can only compare the evidence for one model in relation to others […]. This means that there is no ‘‘true’’ model—there is only the ‘‘best’’ model from among those models considered (Corcoran et al., 2023, p. 7).

Under this Bayesian perspective, scientific progress is thus a process of iterative cycles of revisions and testing, in which the Bayesian machinery can be employed successively over time to determine which, among the models under examination, require less adjustments and modifications to fit the empirical data. In this scheme, it may then be that one theory’s models perform well enough to eventually dominate the Bayesian race for evidence, thereby staking its claim to become the leading paradigm.

Adopting a Bayesian point of view on adversarial collaboration helps with some of the standard criticisms that philosophers of science close to falsificationism have raised against Bayesianism (see Howson & Urbach, 1989; Rosenkrantz, 1977 for discussions), mostly because Bayesianism is seen as an inductivist approach to scientific progress, rather than a deductive one (but see Gelman & Shalizi, 2013 for a discussion).

The first criticism is that accumulation of evidence for a theory should not justify any credence in the theory, because, according to falsificationism, the logic of science consists in trying to falsify predictions, not in verifying them. In Popperian terms, corroboration of theoretical predictions only tells us that a theory has passed severe tests, but not that the theory is true (hence the Popperian distinction between a theory being *trust*worthy vs. it being *test*worthy).

There are two points in response to this criticism. First, it seems that the very structure of Bayesian adversarial collaboration includes a falsificationist element insofar as theorists should strive to design experiments that disambiguate amongst competing predictions; this puts emphasis on designing experiments that generate data that may support one theory's predictions while simultaneously challenging those of the rival theory. Pairing a Bayesian outlook with adversarial collaborations constructed in this way could then demonstrate some compatibility between the idea that science progresses by trying to falsify theories and the idea that it is scientifically rational to attribute higher degrees of belief to theories that accumulate evidence in their favour.

The second point touches instead on the debate between scientific realism (i.e., the position that our best scientific theories and models are at least approximately true, and therefore we are justified in believing in the existence of the observable entities and phenomena posited by such theories) and antirealism (i.e., the position that denies the correspondence between scientific theories and entities/phenomena in the real world) (Okasha, 2016: Ch. 4)[1]. Scholars that adopt a Bayesian view of confirmation theory in the context of adversarial collaborations might very well agree with Popper that a probabilistic increase in the credence for a certain theory does not amount to claiming that the theory is true. Some could endorse an instrumentalist view according to which the model that fits the data best is simply the most useful model, without committing to any claim about its truth. Others, instead, might endorse a structural realist view according to which corroborated scientific theories, despite not being strictly speaking true, are nonetheless able to track some aspects of the structure of reality, and that is enough to underwrite their explanatory and predictive power. Thus, despite there being avenues for linking Bayesianism with realism (Lipton, 2004); a Bayesian need not be committed to the idea that confirmation of a theory depends uniquely on its truth.

A second falsificationist criticism comes from Lakatos (1968), who attacked Bayesian conditionalization by pointing out that it is atheoretical and acritical. According to Lakatos, Bayesianism reduces science to statistics, since the unit of analysis at which conditionalization applies is at the level of low-level predictions about parameter values, but nothing is said about how precisely those predictions are connected to higher-level theories and general explanations. Moreover, in Lakatos' view, Bayesianism is acritical in the sense

[1] A specific and influential brand of antirealism is instrumentalism, which contends that scientific theories are just instruments or tools to predict phenomena and systematize observations, and for these reasons they can be more or less practically adequate, but not necessarily accurate.

that it does not prescribe a way to rule out states of affairs, since the best it can do is just to lower the degree of credence in a specific hypothesis:

> In this [Bayesian] method there is no place of honour accorded any more to *theories or laws*. […] The concept of *explanation* […] disappears; though we may retain the term as a manner of speech for those sentences whose instantiations have high confirmation. *Testability* disappears *too,* for there are no potential falsifiers. No state of affairs is ever excluded. The recipe is: guesses, with different and changing degrees of probability, but without criticism. Estimation replaces testing and rejecting (Lakatos, 1968, p. 348 – italics in the original).

A possible reply to the criticism of Bayesianism being atheoretical is that the relationship between low-level hypotheses and high-level theories could be implemented through *hierarchical* Bayesian models, in which higher levels encode general information and background assumptions that constrain the more fine-grained claims and predictions of the levels below (Grim et al., 2022; Henderson, Goodman, Tenenbaum, & Woodward, 2010).

Moreover, Bayesian scholars can reply that Lakatos' view about Bayesianism being acritical is too crude. For example, the iterative nature of theory testing and model comparison (emphasized by Lakatos' methodology of science too), paired with the falsificationist rationale driving adversarial collaborations, seems to prescribe a methodology for gradually reducing our credence in a specific hypothesis that is practically (although perhaps not logically) equivalent to rejecting it.

This discussion reveals that the Bayesian confirmation-theoretic account proposed by Corcoran et al. in the context of adversarial collaborations can successfully address some of the standard criticisms raised against Bayesian epistemology, and could actually subsume some falsificationist tenets within a hierarchical Bayesian framework.

However, this approach still faces another line of criticism that does not derive from the falsificationist tradition, but from the Kuhnian perspective. The issue is that different theories might be *incommensurable,* in the technical Kuhnian sense: different theories operate with different worldviews, methodologies, and conceptual frameworks. This implies that there might not be a standard way to measure their empirical success, since proponents of different theories "live in different worlds". If this is true, the very notion of scientific progress becomes meaningless. As seen above, this problem is particularly pressing in consciousness science, given its pre-paradigmatic stage.

In addressing this problem, Corcoran et al. maintain that ranking models given how well they account for empirical data affords the development of a common evidential currency that can be used to assess the empirical fitness of different models, driven by different conceptual frameworks. Concerns about commensurability could then be sidestepped by evaluating the quality of model fits to empirical data; that is, by comparing the evidence for different models.

Although this seems to partially address the worry about the commensurability of different theories of consciousness (it is worth noting that for Kuhn himself incommensurability did not mean incomparability – see Sankey, 1993), a further epistemological problem remains: it is not clear that the Bayesian approach in itself can determine whether a theory is a theory *of consciousness*. In other words, the Bayesian account provided by Corcoran and colleagues (which can be generalized and used in different fields) explains how to compare different models given their empirical performance, and therefore the account is well-suited to explain the relationship between data and hypotheses. A hierarchical Bayesian account can further include the relationship between hypotheses and general level theories. This means that the Bayesian strategy allows one to compare different hypotheses, models, and theories by bringing them into the same arena, which is built on the concept of model evidence. But this Bayesian approach does not account directly for the relationship between data and phenomena – it is silent on the question of whether a theory is a theory of consciousness or, for example, attention[2]. For this reason, the acceptance of a theory as paradigmatic for consciousness science will likely be guided by extra-evidential factors too. Such factors, which require conceptual work and argumentation, are necessary for consciousness scholars to accept a theory as a theory *of consciousness* (independently of its capacity to accommodate empirical observations).

The metatheoretical strategy

[2] This is not to say that there are no Bayesian approaches to trace back phenomena from data, but simply to say that this relationship does not seem to be directly accounted for by a confirmation-theoretic approach like the one proposed by Corcoran et al. (2023).

Perhaps, these extra-evidential reasons for choosing one theory over another can be found by analysing how theories are constructed, and the fundamental theoretical and empirical concepts that constitute them. Uncovering these foundational aspects of theories of consciousness is what the metatheoretical strategy attempts to do: according to this strategy, progress in consciousness science is expected to be made by adopting a high-level view on consciousness science and its theories, and more specifically by comparing the theoretical ground of different theories in the neuroscience of consciousness, prior to their empirical comparisons.

This metatheoretical strategy can be implemented in many different ways. The first option is to try to analyse the field of consciousness science to see how theories influence its practice (and vice versa). Examples of this approach include Yaron et al. (2022); Michel et al. (2018), and Francken et al. (2022).

A second way of implementing the metatheoretical strategy is to single out the explanatory constructs and strategies of theories of consciousness, to inform more specific theory-comparison. The basic idea is that if we could find some explanatory dimensions that are common between theories, then theory-comparison would just need to be performed along these dimensions – that is, theories would differ in virtue of how they move along these dimensions. Examples of this approach include Signorelli et al. (2021); Fazekas et al. (2024); Doerig et al. (2021); Evers et al. (2024); Seth & Bayne (2022).

A third option is to extract conceptual commonalities in order to reduce the field's theory space. This has been done either by focusing on specific theories (Northoff & Lamme, 2020; Storm et al., 2024) or by suggesting more abstract ways to unify different theories into broader frameworks for consciousness (He, 2023; Singhal & Srinivasan, 2024; Wiese, 2020).

Finally, the metatheoretical strategy can be implemented by devising protocols and methodologies for empirical theory-comparison. The basic idea, in this case, is that having a clear notion of the epistemic limits of empirical theory-testing can inform experimentation itself, and therefore allow for more structured and meaningful empirical investigation. Examples of this metatheoretical stance are: (Chis-Ciure, Melloni, & Northoff, 2024; Del Pin, Skóra, Sandberg, Overgaard, & Wierzchoń, 2021; Kirkeby-Hinrup, 2024b; Kirkeby-Hinrup & Fazekas, 2021; Negro, 2024)

At this point, it is possible to see why the empirical, confirmational, and metatheoretical strategies should be interpreted as complementary approaches in consciousness science. Although they share the same final goal, which is to make progress in the scientific study of consciousness, they have different strengths and we believe that genuine progress can be achieved only by combining their mutually supporting capacities. In fact, focussing only on the empirical strategy would risk accumulating empirical data without a systematic framework to compare how different theories are affected by experimental results; addressing only the confirmational strategy would risk selecting a best model that is not necessarily a model *of consciousness*; and working exclusively on the metatheoretical strategy would risk obliviating the centrality of empirical data and evidence accumulation for scientific progress.

Although it is not required of every single research programme in consciousness science to include all three strategies, it is beneficial that they be represented in the field at large, so as to ensure a type of progress that is sensitive to both the theoretical and empirical aspects of scientific understanding.

Conclusion

In this chapter we have reviewed several topics in the philosophy of science that can be translated into the science of consciousness, to better understand the epistemic dynamics that move the field forward. In the first section, we argued that a stance on the distinction between data and phenomena can help consciousness scientists define the explanandum and the structure of the explanans; in section 2, we argued that the distinction between models and theories has ramifications for how theories of consciousness are empirically tested, and therefore philosophical analysis on the structural relationship between these concepts bears substantial scientific impact; in section 3, we turned to the question of what makes a theory scientific, and argued that this is an important question for a young field in which theoretical speculations and philosophical theorizing are widespread. In fact, after surveying some of the most popular positions on the demarcation problem, we have suggested that the very question of whether a theory of consciousness is scientific or not might be premature, given the pre-paradigmatic state of the field. Because of this, in section 4 we considered several strategies for making progress in consciousness science.

We believe that these considerations are not only theoretically important, but are also practically significant, since a better grasp on the state of consciousness science as a scientific field – and on its theories as genuine scientific research programmes – can inform the actual practice of science as a societal endeavour. For example, arguments on the legitimacy of theories in institutional contexts (e.g., scientific journals, conferences, etc.) and on the most appropriate distribution of funding and resources can, and should, be informed by the type of considerations from philosophy of science that we have presented here.

Applying a philosophy of science lens to consciousness science thus allows us to see that consciousness scientists are currently facing an exploration-exploitation trade-off, according to which scientists must find the optimal balance between an exploratory strategy (which is epistemically uncertain) and an exploitative one (which takes advantage of a fairly secure methodology). We believe that the conceptual tools from the philosophy of science introduced in this chapter can aid consciousness scientists in navigating this exploratory stage of the field, and help them determine the conditions of possibility for the establishment of a paradigmatic science of consciousness.

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., . . . Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. *PLOS Computational Biology, 19*(10), e1011465. doi:10.1371/journal.pcbi.1011465

Baars, B. J. (2009). History of Consciousness Science. In W. P. Banks (Ed.), *Encyclopedia of Consciousness* (pp. 329-338). Oxford: Academic Press.

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

Bird, A. (2007). What Is Scientific Progress? *Noûs, 41*(1), 64-89. Retrieved from http://www.jstor.org/stable/4494519

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18(2)*, 227-247.

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

Bogen, J., & Woodward, J. (1988). Saving the phenomena. *Philosophical Review, 97*(3), 303-352.

Buccella, A., Maoz, U., & Mudrik, L. (2024). Towards an interdisciplinary "science of the mind": A call for enhanced collaboration between philosophy and neuroscience. *Eur J Neurosci*. doi:10.1111/ejn.16451

Carnap, R. (1956). The Methodological Character of Theoretical Concepts. In H. Feigl & M. Scriven (Eds.), *The Foundations of Science and the Concepts of Psychology and Psychoanalysis* (pp. 38--76): University of Minnesota Press.

Cartwright, N. (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Chalmers, D. J. (2004). How can we construct a science of consciousness? In M. S. Gazzaniga (Ed.), *The Cognitive Neurosciences III* (pp. 1111--1119): MIT Press.

Chis-Ciure, R., Melloni, L., & Northoff, G. (2024). A measure centrality index for systematic empirical comparison of consciousness theories. *Neurosci Biobehav Rev, 161*, 105670. doi:10.1016/j.neubiorev.2024.105670

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

Clark, C. J., & Tetlock, P. E. (2023). Adversarial Collaboration: The Next Science Reform. In C. L. Frisby, R. E. Redding, W. T. O'Donohue, & S. O. Lilienfeld (Eds.), *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions* (pp. 905-927). Cham: Springer International Publishing.

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

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences, 15*(8), 358-364. doi:https://doi.org/10.1016/j.tics.2011.06.008

Corcoran, A. W., Hohwy, J., & Friston, K. J. (2023). Accelerating scientific progress through Bayesian adversarial collaboration. *Neuron, 111*(22), 3505-3516. doi:10.1016/j.neuron.2023.08.027

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

Craver, C. F. (2002). Structures of Scientific Theories. In P. Machamer & M. Silberstein (Eds.), *The Blackwell Guide to the Philosophy of Science* (pp. 55–79): Blackwell.

Crick, F., & Koch, C. (1990). Toward a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2,* 263-275.

Crick, F., & Koch, C. (1998). Consciousness and neuroscience. *Cereb Cortex, 8*(2), 97-107. doi:10.1093/cercor/8.2.97

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

Dehaene, S., Sergent, C., & Changeux, J.-P. (2003). A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences, 100*(14), 8520-8525. doi:10.1073/pnas.1332574100

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

Dennett, D. C. (1995). The path not taken. *Behavioral and Brain Sciences, 18*(2), 252-253.

Dennett, D. C. (2016). Illusionism as the Obvious Default Theory of Consciousness. *Journal of Consciousness Studies, 23*(11-12), 65-72.

Doerig, A., Schurger, A., & Herzog, M. H. (2021). Hard criteria for empirical theories of consciousness. *Cognitive Neuroscience, 12*(2), 41-62. doi:10.1080/17588928.2020.1772214

Doerig, A., Schurger, A., Hess, K., & Herzog, M. H. (2019). The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness. *Consciousness and Cognition, 72*, 49-59. doi:https://doi.org/10.1016/j.concog.2019.04.002

Einstein, A. (1944). *Letter to R. A. Thornton*.

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

Evers, K., Farisco, M., & Pennartz, C. M. A. (2024). Assessing the commensurability of theories of consciousness: On the usefulness of common denominators in differentiating, integrating and testing hypotheses. *Consciousness and Cognition, 119*, 103668. doi:https://doi.org/10.1016/j.concog.2024.103668

Fazekas, P., Cleeremans, A., & Overgaard, M. (2024). A construct-first approach to consciousness science. *Neuroscience & Biobehavioral Reviews, 156*, 105480. doi:https://doi.org/10.1016/j.neubiorev.2023.105480

Feyerabend, P. (1988). *Against method* (Vol. 87): New Left Books.

Forster, M., & Sober, E. (1994). How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science, 45*(1), 1-35. doi:10.1093/bjps/45.1.1

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

Frankish, K. (2016). Illusionism as a Theory of Consciousness. *Journal of Consciousness Studies, 23*(11-12), 11-39. Retrieved from https://www.ingentaconnect.com/content/imp/jcs/2016/00000023/f0020011/art00002

Frigg, R. (2022). *Models and theories: a philosophical inquiry*: Routledge/Taylor & Francis Group.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *Br J Math Stat Psychol, 66*(1), 8-38. doi:10.1111/j.2044-8317.2011.02037.x

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy, 21*(5), 725-740. doi:10.1007/s10539-006-9054-6

Grim, P., Seidl, F., McNamara, C., Rago, H. E., Astor, I. N., Diaso, C., & Ryner, P. (2022). Scientific Theories as Bayesian Nets: Structure and Evidence Sensitivity. *Philosophy of Science, 89*(1), 42-69.

Hanson, J. R., & Walker, S. I. (2021). Formalizing falsification for theories of consciousness across computational hierarchies. *Neuroscience of Consciousness, 2021*(2). doi:10.1093/nc/niab014

Hanson, N. R. (1965). *Patterns of Discovery: An Inquiry Into the Conceptual Foundations of Science*: Cambridge University Press.

Hansson, S. O. (2006). Falsificationism Falsified. *Foundations of Science, 11*(3), 275-286. doi:10.1007/s10699-004-5922-1

He, B. J. (2023). Towards a pluralistic neurobiological understanding of consciousness. *Trends Cogn Sci, 27*(5), 420-432. doi:10.1016/j.tics.2023.02.001

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

Herzog, M. H., Schurger, A., & Doerig, A. (2022). First-person experience cannot rescue causal structure theories from the unfolding argument. *Conscious Cogn, 98*, 103261. doi:10.1016/j.concog.2021.103261

Hesse, M. B. (1966). *Models and Analogies in Science*: Ind.

Hohwy, J., & Seth, A. K. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences, 1*(II). doi:10.33735/phimisci.2020.II.64

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

Irvine, E. (2012). *Consciousness as a scientific concept: a philosophy of science perspective*: Springer.

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

Kahneman, D. (2003). Experiences of collaborative research. *Am Psychol, 58*(9), 723-730. doi:10.1037/0003-066x.58.9.723

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

Kirkeby-Hinrup, A. (2024a). Interdisciplinary Consciousness Studies needs Philosophers of Science. *Filosofiska Notiser, 11*(1), 3-18.

Kirkeby-Hinrup, A. (2024b). Quantifying empirical support for theories of consciousness: a tentative methodological framework. *Front Psychol, 15*, 1341430. doi:10.3389/fpsyg.2024.1341430

Kirkeby-Hinrup, A., & Fazekas, P. (2021). Consciousness and inference to the best explanation: Compiling empirical evidence supporting the access-phenomenal distinction and the overflow hypothesis. *Consciousness and Cognition, 94*, 103173. doi:https://doi.org/10.1016/j.concog.2021.103173

Kleiner, J. (2020). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition, 85*, 102981. doi:https://doi.org/10.1016/j.concog.2020.102981

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness, 2021*(1). doi:10.1093/nc/niab001

Koivisto, M., Salminen-Vaparanta, N., Grassini, S., & Revonsuo, A. (2016). Subjective visual awareness emerges prior to P3. *European Journal of Neuroscience, 43*(12), 1601-1611. doi:https://doi.org/10.1111/ejn.13264

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions* (Second ed.): University of Chicago Press.

Lakatos, I. (1968). 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. (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.

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

Laplane, L., Mantovani, P., Adolphs, R., Chang, H., Mantovani, A., McFall-Ngai, M., . . . Pradeu, T. (2019). Why science needs philosophy. *Proceedings of the National Academy of Sciences, 116*(10), 3948-3952. doi:doi:10.1073/pnas.1900357116

Laudan, L. (1977). *Progress and its Problems: Toward a Theory of Scientific Growth* (Vol. 87): University of California Press.

Laudan, L. (1983). The Demise of the Demarcation Problem. In R. S. Cohen & L. Laudan (Eds.), *Physics, Philosophy and Psychoanalysis: Essays in Honour of Adolf Grünbaum* (pp. 111-127). Dordrecht: Springer Netherlands.

LeDoux, J. E., Michel, M., & Lau, H. (2020). A little history goes a long way toward understanding why we study consciousness the way we do today. *Proc Natl Acad Sci U S A, 117*(13), 6976-6984. doi:10.1073/pnas.1921623117

Lipton, P. (2004). *Inference to the Best Explanation*: Routledge.

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

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

McMullin, E. (2013). The Virtues of a Good Theory. In *The Routledge Companion to Philosophy of Science*: Routledge.

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

Merker, B., Williford, K., & Rudrauf, D. (2021). The Integrated Information Theory of consciousness: A case of mistaken identity. *Behavioral and Brain Sciences*, 1-72. doi:10.1017/S0140525X21000881

Michel, M. (2019). Consciousness Science Underdetermined: A short history of endless debates. *Ergo: An Open Access Journal of Philosophy, 6*.

Michel, M., Fleming, S. M., Lau, H., Lee, A. L. F., Martinez-Conde, S., Passingham, R. E., . . . Liu, K. (2018). An Informal Internet Survey on the Current State of Consciousness Science. *Frontiers in Psychology, 9*. doi:10.3389/fpsyg.2018.02134

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. *Philos Trans R Soc Lond B Biol Sci, 373*(1755). doi:10.1098/rstb.2017.0357

Naccache, L., & Dehaene, S. (2007). Reportability and illusions of phenomenality in the light of the global neuronal workspace model. *Behavioral and Brain Sciences, 30*(5-6), 518-520. doi:10.1017/S0140525X07002993

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*: Harcourt, Brace & World.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review, 83*(4), 435-450. doi:10.2307/2183914

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences, 19*(5), 979-996. doi:10.1007/s11097-020-09681-3

Negro, N. (2022a). Axioms and postulates: Finding the right match through logical inference. *Behavioral and Brain Sciences, 45*.

Negro, N. (2022b). Can the Integrated Information Theory Explain Consciousness from Consciousness Itself? *Review of Philosophy and Psychology*. doi:10.1007/s13164-022-00653-x

Negro, N. (2024). (Dis)confirming theories of consciousness and their predictions: towards a Lakatosian consciousness science. *Neuroscience of Consciousness, 2024*(1). doi:10.1093/nc/niae012

Northoff, G., & Lamme, V. (2020). Neural signs and mechanisms of consciousness: Is there a potential convergence of theories of consciousness in sight? *Neurosci Biobehav Rev, 118,* 568-587. doi:10.1016/j.neubiorev.2020.07.019

Okasha, S. (2016). *Philosophy of Science: Very Short Introduction*: Oxford University Press.

Pitts, M. A., Metzler, S., & Hillyard, S. A. (2014). Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. *Front Psychol, 5,* 1078. doi:10.3389/fpsyg.2014.01078

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.

Rosenkrantz, R. D. (1977). Bayes and Popper. In *Inference, Method and Decision: Towards a Bayesian Philosophy of Science* (pp. 118-134). Dordrecht: Springer Netherlands.

Sankey, H. (1993). Kuhn's Changing Concept of Incommensurability. *The British Journal for the Philosophy of Science, 44*(4), 759-774. Retrieved from http://www.jstor.org/stable/688043

Searle, J. R. (1998). How to study consciousness scientifically. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353*(1377), 1935-1942. doi:doi:10.1098/rstb.1998.0346

Searle, J. R. (2000). Consciousness. *Annual Review of Neuroscience, 23,* 557-578. doi:10.1146/annurev.neuro.23.1.557

Searle, J. R. (2017). Biological Naturalism. In *The Blackwell Companion to Consciousness* (pp. 327-336).

Seth, A. K. (2021). *Being You: A New Science of Consciousness*: Penguin Publishing Group.

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

Shan, Y. (2019). A New Functional Approach to Scientific Progress. *Philosophy of Science, 86*(4), 739-758. doi:10.1086/704980

Signorelli, C. M., Cea, I., & Prentner, R. (2024). We need to explain subjective experience, but its explanation may not be mechanistic. *PsyArXiv*. doi:https://osf.io/preprints/psyarxiv/e6kdg

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

Singhal, I., & Srinivasan, N. (2024). Just one moment: Unifying theories of consciousness based on a phenomenological “now” and temporal hierarchy. *Psychology of Consciousness: Theory, Research, and Practice*, No Pagination Specified-No Pagination Specified. doi:10.1037/cns0000393

Storm, J. F., Klink, P. C., Aru, J., Senn, W., Goebel, R., Pigorini, A., . . . Pennartz, C. M. A. (2024). An integrative, multiscale view on neural theories of consciousness. *Neuron, 112*(10), 1531-1552. doi:https://doi.org/10.1016/j.neuron.2024.02.004

Suppes, P. (1969). Models of Data. In P. Suppes (Ed.), *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969* (pp. 24-35). Dordrecht: Springer Netherlands.

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

Tsuchiya, N., Andrillon, T., & Haun, A. M. (2020). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a science of causal structure theories of consciousness. *Consciousness and Cognition, 79*, 102877. doi:https://doi.org/10.1016/j.concog.2020.102877

Usher, M. (2021). Refuting the unfolding-argument on the irrelevance of causal structure to consciousness. *Consciousness and Cognition, 95*, 103212. doi:https://doi.org/10.1016/j.concog.2021.103212

Usher, M., Negro, N., Jacobson, H., & Tsuchiya, N. (2023). When philosophical nuance matters: safeguarding consciousness research from restrictive assumptions. *Frontiers in Psychology, 14*. doi:10.3389/fpsyg.2023.1306023

Varela, F. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*(4), 330-349.

Vorms, M. (2018). Theories and Models. In A. Baberousse, D. Bonnay, & M. Cozic (Eds.), *The Philosophy of Science: A Companion* (pp. 0): Oxford University Press.

Whyte, C., Corcoran, A. W., Robinson, J., Smith, R., Moran, R., Parr, T., . . . Hohwy, J. (in preparation). The minimal theory of consciousness implicit in active inference.

Wiese, W. (2020). The science of consciousness does not need another theory, it needs a minimal unifying model. *Neuroscience of Consciousness, 2020*(1). doi:10.1093/nc/niaa013

Yaron, I., Melloni, L., Pitts, M., & Mudrik, L. (2022). The ConTraSt database for analysing and comparing empirical studies of consciousness theories. *Nature Human Behaviour, 6*(4), 593-604. doi:10.1038/s41562-021-01284-5

Zeman, A. (2001). Consciousness. *Brain, 124*(Pt 7), 1263-1289. doi:10.1093/brain/124.7.1263