Mathematical Approaches in the Scientific Study of Consciousness

Johannes Kleiner

Preprint, submitted to Melloni, L. & Olcese, U. (Eds.). (Forthcoming). *The Scientific Study of Consciousness – Experimental and Theoretical Approaches*. Springer Nature.
August 9, 2024.

It is perhaps somewhat surprising, from an external point of view, that the scientific study of consciousness, also known as consciousness science, comprises a host of mathematical models, methods, and questions. This is the case, on the one hand, because it builds on mathematical models and mathematical methods developed in other sciences, for example models of the brain, analysis techniques, modelling procedures, or statistical tests. And it is the case, on the other hand, because consciousness itself is amenable to mathematical description and mathematical representation. As a result, the study and exploration of mathematical topics has become a notable task in consciousness science. It is now known as *Mathematical Consciousness Science*.

Mathematical consciousness science is to consciousness science what mathematical physics is to physics, what mathematical biology is to biology, and what mathematical neuroscience is to neuroscience. It is the application and study of formal and mathematical methods as applied in, or relevant to, the scientific study of consciousness. Because mathematical methods and mathematical questions appear in experimental, theoretical, conceptual, and methodological domains in consciousness science, mathematical consciousness science comprises experimental, theoretical, conceptual, and methodological questions.

This chapter aims to introduce the reader to research in mathematical consciousness science in these four domains, so as to provide them with one initial perspective of what mathematical consciousness science is, and how it contributes to the sci-

Johannes Kleiner
Munich Center for Mathematical Philosophy, LMU Munich, Geschwister-Scholl-Platz 1, D-80539 München, Germany & *Graduate School of Systemic Neurosciences, LMU Munich*, Großhaderner Str. 2, D-82152 Planegg-Martinsried, Germany & *Institute for Psychology, University of Bamberg*, Markusplatz 3, D-96047 Bamberg, Germany, e-mail: johannes.kleiner@uni-bamberg.de

entific study of consciousness at large. It aims to do so by reviewing some of the recent progress in mathematical consciousness science in these domains. But it also attempts to provide an outlook into the future, by painting a picture, with broad strokes, of how some of the research in mathematical consciousness science could be further developed, and how it might factor into the future development of consciousness science. The hope is to provide the reader with an impression of the goals and long-term visions that motivate, or have emerged in, parts of mathematical consciousness science.

Due to reasons of space, this chapter is not a review or comprehensive assessment of mathematical consciousness science. The selection of topics is highly biased around the research of the author, and it only mentions a small part of the research that 200+ researchers, who have joined the mathematical consciousness science community as represented by the *Association for Mathematical Consciousness Science*, have produced.

If anything, the term mathematical consciousness science should be taken to refer to the questions which members of this community of researchers pose and answer. Much like it was impossible to foresee the developments of modern mathematical physics prior to the 20$^{\text{th}}$ century, it is impossible to foresee what mathematical consciousness science might become. The following introduction to some of the contemporary themes of mathematical consciousness science should therefore be understood as outlining a very preliminary picture. More comprehensive introductions to some of the research mentioned here can be found in the other chapters in this section of the present volume.

1 Research on Theories of Consciousness

Theories of consciousness, also called models of consciousness, are hypotheses about how conscious experiences and the subject matter of the natural sciences, most notably the brain, relate. They are usually required to be substantive and non-trivial, and are either derived from experimental data or meaningful conceptual assumptions. This section provides an introduction to what mathematical consciousness science can contribute to the research on theories of consciousness.

1.1 How Do We Build Theories of Consciousness?

Consciousness science, in its present stage of development, comprises a large number of theories. There are at least 39 scientific theories of consciousness published in journals that relate to the field,[1] and likely many more in journals of different disciplines or unpublished at the present stage.

[1] This number goes back to a list kindly complied by Dr. Jonathan Mason, quoted in (Kleiner, 2024c). If metaphysical theories are included, the count is much higher even (R. L. Kuhn, 2024).

While some variation among theories might be expected given the different metaphysical assumptions and explanatory strategies that are employed (Signorelli, Szczotka, & Prentner, 2021), the bulk of the variation, arguably, may be due to the fact that there are no noteworthy constraints in proposing a theory of consciousness. Any hypothesis or experimental finding can, once singled out from its context, be presented as a new theory of consciousness. All that is required is the individuation of some property, mode, mechanism or configuration among the subject matter of the natural sciences, as well as some conjecture of how this property, mode, mechanism or configuration might relate to the conscious perception of a stimulus, the instantiation of a phenomenal property, or the subject being conscious at all.

This situation may be due to the fact that consciousness science does not yet have a thorough paradigm, in the sense of T. S. Kuhn (1962), for how to *build* a theory of consciousness. Perhaps the theories that exist are, as far as the construction of theories of consciousness is concerned, more akin to examples. They embed a huge amount of very valuable insights, but a thorough paradigm that guides the construction of theories, like Newtonian mechanics in physics, for example, is not yet available.[2]

A substantial part of the work on theories of consciousness in mathematical consciousness science is aimed at gauging the direction in which such a paradigm for the theory-building process might eventually be found. Here are three examples.

1.1.1 The Role of Mathematics in Constructing Theories of Consciousness

Perhaps the most important question regarding theories of consciousness in mathematical consciousness science is the question of which role mathematics can play in the theory-building process. This includes questions like:

(a) What can mathematics add to the theory-building process in consciousness science? Are there particular advantages when using mathematics to formulate theories of consciousness? And if so, how do these advantages pan out in practice?
(b) Are there reasons for why it might be necessary to use mathematics in formulating theories of consciousness, under certain circumstances?
(c) How do we actually use mathematics to build theories of consciousness?

It is important to stress that the use of mathematical formalism in constructing theories of consciousness is not obviously a good idea. The large majority of theories of consciousness that exist today is not formulated in mathematical terms, and the science is making good progress nevertheless. Any call for mathematization, either in general or in specific cases, needs to introduce good reasons for why a mathematization should be applied. Good reasons that justify not only the effort of applying the mathematization, but which also take into account the further scientific progress of theories. The exactness and devotion to detail involved in mathematical methods can easily inhibit a progressive research programme.

[2] I am very grateful to Tim Ludwig for discussions on this topic.

The following answers to the above questions are an attempt to summarise the motivation of research on theories of consciousness in mathematical consciousness science, but they might be highly biased by the author's own work, particularly work that is directed at uncovering or reconstructing the mathematical structure of existing theories of consciousness, including both *formal* theories, like Integrated Information Theory (IIT) in (Kleiner & Tull, 2021; Tull & Kleiner, 2021) and Predictive Processing Theory with Active Inference in (Tull, Kleiner, & Smithe, 2023), and *non-formal* theories, such as Global Neuronal Workspace Theory (GNWT) and Higher Order Thought Theories (HOT).

A more faithful account of the target phenomenon. Perhaps the most important contribution that mathematics can make to the theory-building process in consciousness science is the use of mathematical structures and mathematical spaces to represent the target phenomenon (usually conceptualised as phenomenal character). We will postpone discussions of details of this contribution to Section 3.1. What is important here is that this affords a more faithful representation of the target phenomenon of consciousness science in theories of consciousness.

The introduction of mathematical spaces or structures requires one to use formal theories of consciousness: *structural theories of consciousness* that provide a hypothesis about how the subject matter of other natural sciences (neuroscience, biology, etc.) relates to a structural representation of conscious experience.

Flexibility and parsimony in explicating ideas, results, or metaphysics. A second contribution which mathematics might bring to the theory-building process in consciousness science is the explication of ideas, experimental findings, or metaphysics in constructing theories of consciousness.

The mathematical method is known for its dedication to detail. But for theories of consciousness, at least in this early state of development of consciousness science, detail may not be what matters most.

The features of mathematics that seem to be of advantage at the present stage of development of consciousness science are, rather, flexibility and precision. Mathematics is flexible because it allows one to define concepts precisely as intended, free of terms that have several meanings or unintended connotations. And mathematics is precise because a mathematical definition requires one to spell out all details that matter. Because of flexibility and precision, mathematics enables one to explicate ideas, experimental findings, or metaphysics in exactly the way one would like to state them, staying true to what the idea, empirical finding or metaphysical assumption actually comprises. This is, in many cases, a large advantage over non-formal language.

Mathematics in the natural sciences. A third reason for why mathematics might, at least at some point of development, contribute to the formulation of theories of consciousness is that mathematics is also the language used in most theories of natural science.

Theories of consciousness are hypothesis about how the subject matter of natural sciences (usually simply called "the physical") and consciousness relate. Therefore, theories of consciousness need to refer to the subject matter of natural sciences in

some form. At the present stage of development, this reference is usually made by use of names of ROIs or neuroanatomical orientations. But the parts of the brain that are referred to by these names or orientations are distinguished in virtue of their structure, function, or processing. And it is at least likely that a detailed account of such structure, function, or processing requires the use of formal terms or formal models.

More generally, theories of consciousness might need to engage with mathematical formalism because:

1. they address brain functions or neural dynamics on a level of precision that is not amenable to non-formal descriptions,
2. they address the subject matter of the sciences on scales other than the brain, or
3. because they rely on principles which cannot be precisely expressed in terms of natural or near-natural language.

An important example of where developments in neuroscience might necessitate the use of mathematics in formulating theories of consciousness science concerns Predictive Processing Theory (PP) and Active Inference. Because PP is a formal theory, Seth and Hohwy (2021)'s call for replacing the search for ROI-based Neural Correlates of Consciousness by a search "through the lens of [P]redictive [P]rocessing" (ibid., p.1), as well as PP-based Computational Phenomenology as proposed in (Ramstead et al., 2022), arguably require theories of consciousness to be formulated in mathematical terms as well. We will discuss the case of Computational Phenomenology in a little more detail in Section 3.2.

Universality. A forth potential contribution derives from a desideratum for theories of consciousness which has recently been proposed by Kanai and Fujisawa (2024): that theories should be formulated in such a way that they can "determine whether a given dynamical system is conscious, irrespective of its origin or composition (e.g. whether it is a biological brain, hurricane or computer)" (Kanai & Fujisawa, 2024). This desideratum is called *universality*.

While very plausible at first look, the desideratum is not uncontested. Researchers in consciousness science whose work is more aligned with methods and theories of cognitive science argue that theories in cognitive science are not universal in this sense. These theories explain phenomena, but target only the brain. Research on the various forms of memory, say, is not required to produce theories that also hold true for AI systems. So why should a theory of consciousness?

This is certainly a fair point. For all we know, consciousness could be instantiated only in biological systems. Or perhaps, if it is instantiated in artificial systems, it is instantiated in an entirely different way.

On the other hand, one could argue that most theories of consciousness intend to say something that is true of consciousness per se, not just consciousness of systems in the scope of neuroscience. These theories seem to want to provide a universal understanding. The same holds true for views that take consciousness to be a natural kind.

Independently of whether one endorses the desideratum or not, it is arguably the case that the desideratum can only be met by theories of consciousness which

are formulated in mathematical terms. This is the case because, if any language is capable of being applied to the vast range of systems that the desideratum requires, it is the language of mathematics. Any adequate description of Large Language Models (LLMs), for example, is a description in terms of mathematics: LLMs are defined in terms of formal models or computer code. And hurricanes afford formal models as well. Hence any universal theory of consciousness will likely have to be a mathematical theory of consciousness.

Unification. Mathematics could, arguably, be required to unify the different theories of consciousness that exist to date.

A wonderful example of this is the model of consciousness developed by M. Blum and Blum (2021), called *Conscious Turing Machine* (CTM), reviewed in Chapter 5.02 of this volume. This is a model of consciousness, where 'model' is understood in the sense of computer science, which is different from the notion of 'model' in neuroscience and other natural sciences. The CTM is meant to fulfil the same role in consciousness science that Turing's model of computation or Shannon's modelling of information fulfilled in computer science. A big part of the work that flows into the definition and exposition of this model rests on incorporating the core formal ideas of other theories of consciousness in that model, ranging from Global Neuronal Workspace Theory (GNWT) and Predictive Processing Theory (PP) to theories about the evolutionary origin of consciousness, for example the proposals by Humphrey (2023).

Transcending Language. A final opportunity of mathematical approaches in building theories of consciousness that should arguably be mentioned is that they can help to overcome the confines of non-formal language. Mathematics might allow to construct theories of consciousness that rely on concepts which cannot properly be expressed in non-formal language, not because of the practical limitations of non-formal language, but because of the logical context in which non-formal language is embedded.

This feature of mathematics might be helpful in addressing minimal phenomenal experiences (Metzinger, 2024), several aspects of which are difficult to express in ordinary language. And it is important in the context of investigations that aim at formulating theories of consciousness based on idealist or dual-aspect metaphysics, because important concepts in some of these metaphysical frameworks are expressed as what appears to be, from a classical logic perspective, contradictory or nonsensical, but nevertheless expresses substantive points. Examples of such models are (Signorelli, Wang, & Khan, 2021; Hoffman & Prakash, 2014; Atmanspacher & Rickles, 2022).

Because mathematical formalism is not necessarily tied to classical logic, or other presumptions and categories of a mind selected for in natural evolution, mathematics offers a natural and arguably unique starting point to construct theories of consciousness that embed, rely on, or are inspired by concepts that transcend non-formal language. Mathematics, in a sense, allows one to "climb out of" the edifice of a mind selected for by natural selection. Next to precision and flexibility, this may

be a main reason behind the success of mathematics in foundations of physics and related disciplines.

This concludes a very initial assessment of how mathematics might be of help—or, in some cases, even necessary—in constructing theories of consciousness (questions **(a)** and **(b)** above).

Regarding question **(c)**—of how to go about building mathematical theories of consciousness—, the present results suggest only one important answer: there is no one-fits-all solution. Mathematical theories of consciousness are of the same type as non-formal theories of consciousness: they express hypotheses of how the subject matter of the natural sciences relates to conscious experience. The difference lies only in the language that is used to represent the subject matter of the natural sciences, conscious experiences, and the relation that holds between them. But this difference in language does not reduce options. On the contrary, it allows for more freedom in formulating theories of consciousness. Which formulation is appropriate depends on the metaphysics, idea, or empirical finding that motivates the theory. The actual work of explicating this idea will have to be done in each case separately. Making use of the methodological opportunities reviewed above in concrete cases remains challenging and requires as innovative research as always.

It goes without saying that the mathematical effort is complementary to, and not in competition with, non-formal approaches to theory-building. Mathematical consciousness science extends and amends non-formal approaches, in the very same way as theoretical and mathematical physics extend and amend experimental physics. The hope is that a combination of formal and non-formal approaches can ultimately lead to a paradigm in constructing theories akin to the powerful paradigms available in other natural sciences.

1.1.2 Constraints for Theory-Building

A second line of research in mathematical consciousness science that targets the question of how to build theories of consciousness concerns the identification of *constraints* for theories of consciousness, meaning: of conditions that have to be satisfied by the theories.

In what follows, we introduce one such constraint, violation of which is precisely what implies that these theories are "wrong or (...) outside the realm of science", as claimed by the recent *unfolding argument* (Doerig, Schurger, Hess, & Herzog, 2019).

Constraint. The upshot of (Kleiner & Hoel, 2021) and (Kleiner & Hartmann, 2023) is that theories of consciousness have to satisfy the following constraint, for otherwise they cannot actually be empirically investigated:

A theory of consciousness must explain
why measures measure what they measure.

Here, 'measures' refers to measures of consciousness, or C-tests, introduced in more detail in Section 2 both of which are tools to infer information about whether a system is conscious, or what it is conscious of.[3] 'Explain' is a shorthand for the requirement that a theory of consciousness must afford an explanation. Therefore, the constraint states that a theory must explain why measures of consciousness work as intended, if the theory is true, at least for those measures that are used to test the theory's predictions.

Virtually all theories of consciousness that have been proposed to date fail to satisfy this requirement, for one of two reasons. Those theories that offer a specific and well-defined hypothesis about how conscious experience relates to the brain, such as the formal part of IIT, mostly fail because they do not model how subjective report, or other behaviour that measures and C-tests rely on, comes to reflect the content of experience as inferred by the respective measures and C-tests. For example, the specific hypotheses they propose do not contain a concrete explanation of why or how subjects can report on their experience. There is no explanation of how, when a subject reports on their experience, that report can depend on the content of experience according to the theory.

The theories that, on the other hand, operate with a less specific hypothesis about how conscious experience relates to the brain (as is the case, for example, for GNWT, which does not yet specify the necessary or sufficient conditions for something to count as a global workspace), fail for another reason. While the broader picture they propose often includes a sketch of a mechanism that might explain why report or other behavioural indicators come to reflect the content of consciousness, the specifications are too general to afford of an explanation of why measures measure what they measure.

As a result, in both cases, there is a theoretical possibility to vary the part of the brain that is responsible for conscious experiences, according to the theories, without changing the report. Such variations are called substitutions in (Kleiner & Hoel, 2021). In the case of IIT, for example, substitutions of a recurrent network by an "unfolded" feed-forward network as in (Doerig et al., 2019) or a finite automaton by an isomorphic finite automaton as in (Hanson & Walker, 2019) are only special cases of a huge class of substitutions that can be performed. In the case of GNWT, to give another example, one can substitute the part that constitutes the workspace by a simple lookup-table system. Cf. (Kleiner & Hoel, 2021) for more detailed expositions of these examples.

In both cases, the experience changes, but not the behaviour. As explained in detail in (Kleiner & Hoel, 2021), this leads to the problems that Doerig et al. (2019) have first spotted for IIT-like theories.

---

[3] The notion of C-test, proposed in (Bayne et al., 2024), was not available at the time when (Kleiner & Hoel, 2021) and (Kleiner & Hartmann, 2023) were written. But all analyses provided in (Kleiner & Hoel, 2021) and (Kleiner & Hartmann, 2023) apply to both C-tests and measures. That is the case because C-tests have the same formal structure as measures; both are methodologies to infer states of consciousness from experimental data. The difference is which data is considered, and what the states of consciousness that are inferred describe.

As we will see in the next section, implementing this constraint means breaking with some assumptions that are deeply embedded in consciousness science at the present stage of its development. This results in challenges for the field, most notably for theory-building, but also yields novel opportunities, in both theoretical and experimental research.

Before discussing these implications, it should be noted that this is only a first constraint that has been discovered. It is very likely that there are more constraints on how to build theories of consciousness, which are not known at the present stage. Further research of measures and C-tests, perhaps in the form of a measurement theory for consciousness science (cf. Section 2.3), and of consciousness' unique epistemic context is required.

Implementation. Implementing the constraint explained in the last section requires detailed analyses of individual theories of consciousness. Because every theory has its own proposal for how consciousness relates to the brain, it also needs its own explanation, perhaps in the form of a mechanism, of how the content of consciousness comes to determine report in just the right way. To provide an explanation that features enough details to preclude the substitutions explained in the last section is likely no easy feat; a general recipe is probably not available.

There is, however, a minimal condition that all theories have to meet, and which can serve as a starting point for investigations into how individual theories of consciousness could address the constraint. This minimal condition has first been identified in (Kleiner & Hartmann, 2023), in the context of the study of epistemic implications of a particular metaphysical assumption, and was refined in (Kleiner & Ludwig, 2023), so as to present it in a form that underlines its complete metaphysical neutrality. It is now called *dynamical relevance*.

The definition of dynamical relevance, which we will review momentarily, makes use of the fact that scientific theories of consciousness are built on top of the knowledge and insights of natural science: theories of consciousness make use of models or theories from natural sciences to express their hypotheses of what consciousness is and how it relates to the subject matter of the natural sciences. The model or theory from natural science that a theory of consciousness T is built on, or makes use of, can be called T's *reference theory*.

Dynamical relevance, then, in simple terms, is the requirement that a theory of consciousness posit that consciousness makes some difference (= is relevant) to the time evolution of the states of its reference theory. The time evolution of the states of a theory is sometimes called a theory's *dynamics*, hence the name *dynamical relevance*. Cf. (Kleiner & Ludwig, 2023) for a more careful exposition and definition of this notion.

Dynamical relevance is a minimal condition. Any account of how a system's reports or behavioural indicators, as described by the reference theory, come to depend on consciousness as posited by the theory of consciousness, is an account of how consciousness is relevant for the system's dynamics as described by the reference theory. It is important to note that this condition is not in conflict with physicalism, but rather rests on the fact that reference theories—models of the brain, for example—express a particular state of knowledge in the sciences, which the

theory of consciousness amends. Dynamical relevance could be cashed out in terms of a causal influence, but if so, between consciousness as a physical phenomenon and the brain as a physical phenomenon. All that is required is that consciousness get a proper role in the cognitive architecture instantiated by the brain, as far as reports and other behavioural indicators of consciousness are concerned.

Implications. Because dynamical relevance is a minimal condition that must be met in order to resolve the constraint explained in Section 1.1.2, it can serve as a basis to analyse which implications the constraint has for consciousness science at large.

In the last section, we have already alluded to implications for theories of consciousness. At the very least, the constraint requires a fundamental change in theoretical thinking in the context of theories of consciousness. In addition to the "from brain to consciousness" direction, it requires us to engage in the "from consciousness to brain" direction as well, so as to specify which difference consciousness makes in our understanding of brain function.

But it would be wrong to think that these implications merely affect theories. What the constraint, and the analysis in (Kleiner & Hoel, 2021) show, is that we need to step away from conceiving theories and measures as independent from each other. Rather, we must develop a coherent perspective on testing of consciousness, in which both the theory's prediction and the measure's function are considered together. Both are part of understanding, and deriving, predictions in consciousness science, and they cannot be separated.

1.1.3 Structural Theories of Consciousness

A third line of research in mathematical consciousness science that targets the question of how to build theories of consciousness concerns the use of mathematical structure and mathematical spaces in theories of consciousness.

One important advantage of using structure and spaces for building theories of consciousness has already been mentioned in Section 1.1.1: structures and spaces enable theories of consciousness to represent phenomenal character more faithfully. This might constitute a desideratum for theories of consciousness in its own right, but the use of structure also has practical—and quantifiable—consequences that are of independent interest: a larger explanatory scope and an increase, everything else being equal, in predictive power.

Structural theories of consciousness have a larger explanatory scope than their non-structural counterparts, because structures and spaces represent phenomenal character more comprehensively than non-structural approaches can do. They represent the various qualities or phenomenal properties that are instantiated in single experiences, but they also represent phenomenal relations that hold between them. Furthermore, structures and spaces can represent more features and more details of phenomenal character than would otherwise be possible. As a consequence, structural theories can explain more of phenomenal character than their non-structural counterparts are capable of.

Structural theories of consciousness have, everything else being equal, a larger predictive power than their non-structural counterparts, because a structural theory of consciousness constitutes a much more detailed and rigorous hypothesis of how conscious experience relates to the brain.

Consider, for example, Global Neuronal Workspace Theory (GNWT). GNWT does not, in its current formulation, offer an account of how phenomenal character is determined, it only explains how signals from parallel processors enter consciousness. But a promising idea about the relation between GNWT and phenomenal character—that it is the content of the workspace that determines phenomenal character in full—is available in the field. Structural tools allow to turn this idea into a scientific hypothesis, which results in a range of additional predictions that could, in principle, be tested.

The exploration of structural tools for building theories of consciousness has just begun. For some important work see (Kob, 2023; Lyre, 2022; Benjamin & Kob, 2023), among many others. For work devoted on the foundations of the use of structure and spaces in consciousness science, cf. (Lee, 2021; Kleiner & Ludwig, 2024; Kleiner, 2024b).

1.2 Improving & Clarifying Theories

A second major task which mathematical consciousness science has taken up in regard to theories of consciousness is to improve and clarify them.

1.2.1 Integrated Information Theory

Consider, as an example, *Integrated Information Theory* (IIT), introduced in Chapter 2.07 of this volume. IIT is considered one of the leading models of consciousness and aims to describe both the quality and quantity of the conscious experience of a system, such as the brain, in a particular state.

IIT comprises two main parts. On the one hand, a conceptual part that spells out IIT's metaphysical presumptions, as well as a set of assumptions that are taken to characterise the essential properties of conscious experiences in full. The latter are referred to as IIT's 'axioms'. On the other hand, it comprises a complex and rather complicated set of mathematical equations that determine the conscious experience of any system, such as the brain, based on a formal description of the system. This formal part of IIT constitutes the actual hypothesis about how the subject matter of the natural sciences (e.g., the brain) and conscious experiences relate. The conceptual part arguably serves as a justification of the formal part: the formal part is intended to be derived from the axioms and metaphysical premises of the conceptual part.

What mathematical consciousness science can contribute to the development and public understanding of IIT is:

(a) the explication and clarification of which mathematical object it is that the equations and formal concepts of IIT actually describe,
(b) the exploration and assessment of problems of IIT's formal constructions, in particular based on the clarification of the mathematical structure uncovered in **(a)**, and
(c) ways to define the formal content of IIT in terms of more apt mathematics, both to propose improvements of the theory and to make it easier to understand.

Task **(a)** has been carried out as by Kleiner and Tull (2021). The goal of this work was to uncover the mathematical object that underlies the formal descriptions and equations of IIT 3.x, meaning: of any published paper (including supplementary material) that has been published by the lab that develops IIT after the IIT 3.0 paper (Oizumi, Albantakis, & Tononi, 2014) and before IIT 4.0 was first proposed in parts in (Haun & Tononi, 2019). The result of this work is a detailed description and definition of the mathematical mapping that constitutes the formal part of IIT. This mapping maps every formal description of a system, together with a state thereof, to a space of conscious experiences, and element thereof.

A surprising discovery in this respect was that much of the mathematical structure that appears to be essential for IIT's description of conscious experience in terms of formal spaces is actually auxiliary and merely derives from the particular notion of (network-like) classical systems that has been applied in previous expositions of the theory. The mathematical investigation carried out as part of task **(a)** allows to delineate between this essential and auxiliary structure. This matters, for example, for assessments of IIT's phenomenological implications, as well as for theoretical work that attempts to put Global Neuronal Workspace Theory (GNWT) (Dehaene, Changeux, & Naccache, 2011) on a par with IIT as far as explanatory scope is concerned.

Task **(b)**—the exploration of reasons to criticise IIT's formalism—has been a prominent and important part of the literature on IIT since it's full formalism was first proposed by Oizumi et al. (2014). Important early examples are (Barrett, 2014; Cerullo, 2015; Moon & Pae, 2018; Barrett & Mediano, 2019). A particularly important criticism, called 'unfolding argument', was proposed by Doerig et al. (2019), and has inspired investigations of IIT's scope (Michel & Lau, 2020), methodology (Negro, 2020), mathematical framing (Tsuchiya, Andrillon, & Haun, 2019) and testability (Kleiner, 2020a; Ganesh, 2020; Hanson & Walker, 2021). The result of Kleiner and Hoel (2021)'s investigation, which we have already reviewed in Section 1.1.2, shows that there is a fundamental issue with testing IIT that derives both from IIT's mathematical formalism and the typical paradigm of testing theories of consciousness. Importantly, this is not an issue that pertains to IIT alone. Rather, this issue appears for all major theories of consciousness presently proposed, in a nutshell because much like one can, in theory, substitute any recurrent system (which is conscious according to IIT) by a feed-forward system (which isn't conscious according to IIT), while keeping the input-output mapping of the system the same, one can in theory substitute any part of a system (for example, a global workspace) by a

look-up-table device without changing the input-output mapping of the system as a hole.

First steps towards Task **(c)** have been carried out by Tull and Kleiner (2021). The goal of this project was to consider IIT in the context of the powerful and elegant mathematical language of *category theory*. To this end, we have demonstrated how integrated information and other key notions from IIT can be studied within the simple graphical language of process theories (symmetric monoidal categories). As in the work on the mathematical structure of IIT, our desideratum was to stay true to the definitions of IIT 3.x provided in the literature.

The result of this work allows IIT 3.x to be generalised to a broad range of physical theories and sets the foundation for a categorical definition of IIT. A full categorical version of IIT that presents the theory in terms of a functor, however, requires breaking with the formalism of IIT that is published in the literature, and hence is not available to date. The exploration of IIT's relation to category theory, however, is thriving, see for example (Tsuchiya & Saigo, 2021; Tsuchiya, Phillips, & Saigo, 2022; Tsuchiya, Saigo, & Phillips, 2023; Prentner, 2024a), and the summary of research on the structure of qualia in Chapter 5.04 of this volume.

There are many more questions for mathematical consciousness science to consider in relation to IIT, and it is likely that ultimately, IIT can only overcome the various criticisms of its formal structure that have been proposed if it engages with the contributions that are made as part of mathematical consciousness science, most notably those that suggest improvements of the theory. For example, as part of the investigation of the mathematical structure of IIT, Kleiner and Tull (2021) have made a proposal of how IIT's formalism could be amended to overcome the criticism put forward in (Barrett & Mediano, 2019). The amended definition of IIT proposed in this section furthermore is such that the problem discussed in (Hanson & Walker, 2023) cannot occur, qua definition.

1.2.2 Predictive Processing and Active Inference

A second example of improvement and clarification work concerns Predictive Processing Theory (PP) and its Active Inference doctrine, also known as 'Free Energy Principle' (Parr, Pezzulo, & Friston, 2022). While not itself a theory of consciousness, this is arguably a first comprehensive theory of brain function. Because PP and Active Inference aim to offer one coherent principle that explains phenomena as diverse as perception, cognition, planning and action, a connection to conscious experience is not surprising.

While there are comparably simple conceptual ideas that afford a substantial understanding of the theory—prediction, prediction error, prediction error estimation, precision, and so fourth—, the theory is in fact a formal theory of the brain, and only a formal account can grasp the theory in full (Buckley, Kim, McGregor, & Seth, 2017; Parr et al., 2022). What is more, recent expositions of the theory have

moved away from formal structures where concepts like prediction error still play an important role, and towards a formal structure that is independent of, and more general than, these ideas, most notably the 'Factor Graph' formulations (De Vries & Friston, 2017). A mathematical exposition and analysis are therefore helpful not only for inner-theoretic purposes, but also to make the theory accessible for further theorising, in particular where consciousness is concerned.

Tull et al. (2023) provide a categorical formulation of Predictive Processing Theory (PP) with Active Inference, expressed in terms of a simple diagrammatic formal language known as string diagrams that define a monoidal category from the branch of mathematics known as category theory. This research includes diagrammatic accounts of generative models, Bayesian updating, perception, planning, Active Inference, and Free Energy, as well as a diagrammatic derivation of the formula for Active Inference via Free Energy minimisation. As part of this project, we also established compositionality of Free Energy, allowing Free Energy minimisation to be applied at all levels of an agent's generative model. Aside from aiming to provide a helpful graphical language for those familiar with Active Inference, the goal was also to provide a concise formulation and introduction to the Active Inference framework for use in mathematical consciousness science.

Rudrauf et al. (2023) and Sergeant-Perthuis, Rudrauf, Ognibene, and Tisserand (2023) offer an extension of PP and Active Inference that is based on projective geometry. It takes into account important phenomenological observations regarding a first-person perspective, while also generalising the process of active inference to incorporate mathematical group structures that might derive from external or internal spaces of a system.

The hope behind the research carried out in (Tull et al., 2023) is to provide a mathematical basis that allows to formulate hypotheses about how PP and Active Inference relate to conscious experience in concise and mathematically rigorous terms. This is relevant to understanding and clarifying the various hypotheses (cf. for example (Miller, Clark, & Schlicht, 2022)) and methodological ideas (cf. for example (Seth & Hohwy, 2021)) that have been put forward in this context, and constitutes a foundation for future mathematical research on computational phenomenology, cf. Section 3.2.

1.2.3 The Important Case of Non-Formal Theories

Integrated Information Theory and Predictive Processing are, at the present stage of development, of particular interest to mathematical consciousness science because they are the only formal theories within the Overton window of consciousness science at large.[4] But it is important to note that the task of improving and clarifying theories of consciousness also concerns theories which are not presented in mathematical form at the present time.

[4] Several other formal theories of consciousness exist, for example (L. Blum & Blum, 2022) or (Mason, 2021), two mention two.

One reason for an interest in and possible contribution to non-formal theories is that many theories of consciousness employ what could be called 'proto-formal' concepts: concepts that allude or refer to formal notions, but are not presented in a formal form. Another reason is that detailed descriptions of neuronal dynamics and brain functions are formal in nature, and if a theory of consciousness claims that consciousness supervenes on, or is identical to, neuronal dynamics or a brain function, it must take their formal structure into account. Either way, formal ideas, concepts and definition are already part of the theories, albeit mostly not in an explicit way.

Consider, as an example, Global Neuronal Workspace Theory (GNWT), which posits that a system has conscious representations only if two necessary conditions are satisfied. The first necessary condition is that the system has "two main computational spaces, each characterized by a distinct pattern of connectivity" (Dehaene et al., 2011). The first computational space is a "processing network, composed of a set of parallel, distributed and functionally specialized processors or modular subsystems subsumed by topologically distinct (...) domains with highly specific local or medium-range connections" (ibid.); the other computational space is "a global neuronal workspace, consisting of a distributed set of (...) neurons characterized by their ability to receive from and send back to homologous neurons in other (...) areas horizontal projections through long-range excitatory axons" (ibid.), cf. (Kleiner, 2020b) for a more detailed summary and first formal exposition. The second necessary condition is that "[t]he entire workspace is globally interconnected in such a way that only one such conscious representation can be active at any given time" (Dehaene et al., 2011).

This characterisation of the theory is good enough for contemporary purposes and contemporary experimental investigations. But for the theory to properly handle the question of consciousness in organisms and systems that differ from the standard case of healthy humans, the theory must specify which structure, precisely, counts as a computational space of each kind, and what the necessary "patterns of connectivity" are. Computational spaces and patterns of connectivity are formal concepts, hence ultimately a formal specification is in order.

The clarification, improvement or construction of the mathematical structure of existing theories of consciousness is particularly important in the context of Artificial Intelligence (AI), when investigating the possibility of AI consciousness. Because AI systems are formal systems, a rigorous application of theories of consciousness to AI systems cannot do without such formal expositions.

2 Research on Modelling Experiments

Consciousness science is an inherently empirical discipline. Its progress rests on empirical observation in carefully designed experiments that live up to the highest statistical and methodological standards. While mathematical consciousness science

is not concerned with running experiments, it can contribute to the task of designing and abstractly analysing experiments.

2.1 Measures of Consciousness & C-Tests

Because consciousness is not publicly observable "just like that", running an experiment that targets conscious experiences differs substantially from experiments in other sciences: it requires means to infer information about the conscious experience of a subject in experimental trials. Such means are called *measures of consciousness*, cf. e.g. (Seth, Dienes, Cleeremans, Overgaard, & Pessoa, 2008). Simply put, measures of consciousness are "consciousness detection procedures" (Michel, 2023) that can be used to determine whether a subject in an experimental trial has experienced a stimulus consciously or not. There are various measures designed for different paradigms. A simple and effective measure is subjective report: asking a subject whether it has experienced a stimulus consciously or not. But many other measures have been developed as well, in particular to target close-to-threshold stimulus conditions, where subjective reports become unreliable.

A closely related concept are 'C-tests' (Bayne et al., 2024). C-tests are means to infer whether a system is conscious at all, meaning: whether it has conscious experience at all, or not. Like measures of consciousness, C-tests make use of empirical experimental data obtained from individual systems and organisms, but unlike measures, they do not seek to infer information about the particular conscious experience in experimental trials. Rather, they aim to test whether a system has conscious experiences at all. Being able to test whether a system is conscious has a huge clinical importance, and matters largely for ethical, judicial and governance questions.

Measures of consciousness and C-Tests are essential for consciousness science to make progress. That is the case because given the contemporary paradigm of constructing neuroscientific theories of consciousness (more on that paradigm in Section 1.1), they offer the only means to test neuroscientific theories of consciousness in a lab.[5]

Primarily, neither measures nor C-Tests are mathematical in nature. They are based on subjective reports, objective performance or behavioural measures. What mathematical consciousness science can contribute to the study of measures and C-Tests, however, is to model them formally in conjunction with other hypotheses, for example concerning theories of consciousness.

Such an analysis constitutes the foundation of (Kleiner & Hoel, 2021). Because both measures of consciousness and C-tests operate on data obtained in experimental trials, they can both be modelled in terms of a formal mapping

[5] Contemporary neuroscientific theories of consciousness predict the conscious experience that a system has in a particular state. They do not, or are not usually taken to, posit changes in brain dynamics or brain functions. Hence they can only be tested by comparing the prediction the theory makes with the state of consciousness inferred from a measure of consciousness or C-Tests.

$$\inf : \mathcal{O} \to E \,,$$

where $\mathcal{O}$ denotes a set of experimental data sets, and E denotes an abstract notion of states of consciousness, which depending on the application can be either states that target individual experiences, or states that are meant to assess whether as system has consciousness at all, or not. In the case of measures of consciousness, $\mathcal{O}$ comprises data obtained in individual experimental trials, for example in contrastive analysis methods; in the case of C-Tests, $\mathcal{O}$ denotes data obtained in testing a subject of interest, for example behavioural indicators of a subject in the case of a non-responsive wakefulness test; the mapping inf then represents the particular rules and operations that result in 'conscious' vs. 'unconscious' judgements.

What we found in (Kleiner & Hoel, 2021) was that if such consciousness inference procedures are independent from a theory's prediction—a situation which prima facie one would think is ideal—a very counter-intuitive result follows: for every correct inference of an experience, one can modify the part of the system that matters for consciousness so as to obtain a different, non-overlapping prediction while keeping the inferred state constant. Hence the theory must be false, or (if no correct inference exists) untestable. This analysis provides the exact formal underpinning of the unfolding argument that criticises IIT (Doerig et al., 2019). The analysis shows that the argument applies to experimental paradigms, which comprise both theory and measurement, rather than theories alone, and applies to a wide range of theories. More details of this problem, and ways to resolve it, are discussed in Section 1.1.2.

2.2 The Closure Paradigm

A second example of modelling experiments in consciousness science is (Kleiner & Hartmann, 2023). Unlike the research reviewed in the previous section, this work does not model the inference process via measures of consciousness or C-tests. Rather, it models experiments on a more fundamental level, the level of data collection and data storage.

What this work shows is that a central paradigm that spans experimental and theoretical work in consciousness science needs revision: the paradigm that consciousness science is to take a neuroscientific account of the brain as "input", so as to explain what consciousness is, without amending or adding to this input—without amending or adding to the wealth of neuroscientific knowledge, that is. This 'closure paradigm', as one might call it, is at the heart of both identity theories and functionalist theories of consciousness, and is intimately related to discussions of physicalism and the closure of the physical in philosophy of mind.

Kleiner and Hartmann (2023) show that this paradigm conflicts with the testability of theories of consciousness. The underlying intuition is simple: if theories don't amend, or add to, the neuroscientific model of the brain, they cannot account for how experimental data, which relies on reports or behavioural indicators, all of which

are subject to the neuroscientific model, speaks in favour of one, rather than another theory.

This has profound implications for how theories of consciousness should be formulated, cf. Section 1.1. At the very least, consciousness needs to be *dynamically relevant* with respect a reference neuroscientific model, meaning: it must be relevant to the dynamical evolution of a neuroscientific system over and above the dynamical evolution prescribed by the reference model, for otherwise two theories of consciousness cannot disagree about the report that should ensue in experimental trials.

Put in somewhat general terms, this result could be read as a requirement for consciousness to have genuine causal powers,[6] but it is important to note that these need not be extra-physical. On the contrary, it is specifically in the case of physicalist and neuroscientific assumptions that the full force of this result applies: consciousness needs to be understood as a full-fledged physical part or process of the brain; it cannot be subjugated to an epiphenomenon of sorts.

2.3 What is Measurement in Consciousness Science?

Contrary to public conception, consciousness can be measured. There are several proven measures of consciousness and C-tests that empirical investigations in the field make us of.

What there isn't, at present, however, is a substantive *theory of measurement* (or *measurement theory*) for consciousness science that provides a foundation for measurements of the various forms.

Measurement theories are an important part of those sciences where measurement is not straightforward. Consider, as an example, the case of psychology. In the first part of the 20th century, the question of whether there is measurement at all in psychology has been heavily debated, so much so that in 1932 a committee of the British Association for the Advancement of Science was appointed "to decide whether or not there was such a thing as measurement in psychology" (Borsboom, 2005), cf. (Ferguson et al., 1940). The committee's report was highly divided, with a majority of members around the physicist Norman Campbell strongly rejecting claims about the possibility of measurement in psychology.

In response to this rejection by part of the committee, psychologists started to develop theories of measurement that are targeted specifically at psychological experiments, first in the form of scales by Stevens (1946) and then in the form of axiomatic theories of measurement (Cliff, 1992), the most well-known of which is foundational measurement theory (Krantz, Luce, Suppes, & Tversky, 1971), also called representational measurement theory. These developments were pivotal to the

[6] There are many different interpretations of what causal language should mean, cf. for example (Beebee, Hitchcock, Menzies, & Menzies, 2009). This is why the concept of dynamical relevance is formulated without reference to causation.

progress of psychology in the 20th century (Michell, 1999; Borsboom, 2005), and still form the basis of much of the empirical work that is being carried out.

Another example of where a theory of measurement was crucial for understanding the intricacies of measurement is quantum theory in physics. Quantum theory comprises a comprehensive account of measurement, where complicated measurement apparati, that might fill a whole room in a lab, are represented by comparably simple mathematical objects: self-adjoint operators on Hilbert spaces,[7] which in finite-dimensional cases, and given a choice of basis, can be thought of as matrices over complex numbers. Based on these mathematical representations of measurement devices, quantum theory provides an account of how measurement interfaces with the time evolution of a system, which includes an account of possible results of the measurement procedure, as well as an account of how the measurement procedure changes or modifies the state of the system. While this account of measurement is also the source of the notorious measurement problem of quantum theory (cf. e.g. (Myrvold, 2022)), it is hard to imagine which progress could have been possible without the introduction of this part of the theory by von Neumann (1932).

Given consciousness' unique epistemic context, it is likely that a theory of measurement will be as consequential for consciousness science as it has been for psychology and quantum physics. And given that measurement theory in both psychology and physics is based on a mathematical representation of the measurement process, it is at least prima facie possible that mathematics might offer the right tools to develop a measurement theory for consciousness science as well. Whether or not this is possible, or fruitful, remains an open question, and scepticism is fully warranted. But in light of the importance (and, in some cases, necessity) of mathematical theories of measurement in other sciences, the possibility seems worth exploring.

Building on the examples of psychology and physics, a preliminary list of desiderata for a measurement theory of consciousness science could comprise the following elements:

(a) A mathematical representation of measurement. Both psychology and physics, despite being different in many respects, rest on mathematical representations of measurement procedures. Correspondingly, a measurement theory for consciousness science will likely have to comprise mathematical representations of measurements in consciousness science. Such representations should be both descriptive (in the sense that they are built on and represent the actual empirical measurements that are carried out) and normative (in the sense that they guide the development of novel forms of measurement).

(b) What constitutes a measurement, and what not? The mathematical representations of measurement should account for what measurement of consciousness is, and delineate between what counts as measurement, and what does not. For example, does simple verbal report constitute a measurement? And how far can no-report paradigms be pushed while still constituting measures of consciousness?

[7] In present-day quantum information, measurements are represented by the more general concept of *quantum instruments*, developed by Ozawa (1984).

(c) Which aspects of consciousness are measurable, and which not? A theory of measurement for consciousness science should provide conditions that have to be in place for some aspect (or property, mode, part, etc.) of conscious experience to be subjected to measurement.

(d) How does measurement integrate with theories of consciousness? A measurement theory should furthermore account for how measurement integrates with theories of consciousness. In fact, lack of such integration is precisely what plagues the theory-building process in consciousness at the present stage of development, cf. Section 1.1.2.

(e) Which conditions make measurement possible? The abstract representations of measurement should, ideally, also provide conditions that make measurement possible. This is the case, for example, for representational measurement theory in psychology, which relates measurement of a qualitative structure to representation and uniqueness theorems regarding quantitative representations of that structure.

Some progress has already been made in regard to points **(a)** and **(d)**, albeit in unsystematic form. Regarding **(a)**, a first mathematical representation of measures of consciousness, which also applies to C-Tests of consciousness, has been developed in (Kleiner & Hoel, 2021), cf. Section 2.1 This representation has subsequently been applied to investigate problems and necessary conditions of the integration of measures of consciousness in theories of consciousness as required by **(d)**, cf. Section 1.1.2 for a summary. Furthermore, the work on foundations of structural methodologies in (Kleiner, 2024c), (Kleiner & Ludwig, 2024), and (Kleiner, 2024b) might contribute to exploring **(c)** when combined with ideas from axiomatic measurement theory.

It is needless to say that these results constitute only very initial steps, and a further exploration of all of the above-mentioned questions is urgently needed. The hallmark of a successful measurement theory, the allocation of new measurement procedures to a field, as for example the case with additive conjoint measurement (Luce & Tukey, 1964) in psychology, is still nowhere in sight.

3 Research on Conceptual and Methodological Questions

Concepts and methods are essential for any science to move forward. As Daniel Dennett once put it famously, "there is no such thing as philosophy-free science; there is only science whose philosophical baggage is taken on board without examination" (Dennett, 1995, p. 21). This is particularly true of consciousness science, where a large number of concepts have been developed in order to refer to the target phenomenon and make it accessible to scientific analysis. This work is all but finished, and research on new concepts to describe, refer, or represent (parts of) the target phenomenon in consciousness science is a particularly important contribution of philosophy of mind to the science of consciousness.

Good concepts are required to develop rigorous experiments and theories, and are essential to avoid mistakes in theorising (Nida-Rümelin, 2018). Conceptual work is an essential but often overlooked ingredient in pushing the boundaries of scientific knowledge. Correspondingly, a third pillar of how mathematical consciousness science can contribute to the scientific study of consciousness is the exploration and analysis of formal concepts in consciousness science.

3.1 Promises and Foundations of a Structural Turn

A particularly noteworthy development related to formal concepts in consciousness science is the introduction of mathematical spaces and mathematical structures in order to describe or represent conscious experiences as part of the scientific methodology. While pioneering work in this respect has been carried out right in the initial phases of the field (Clark, 1993; Rosenthal, 2010), mathematical spaces and structures are now applied in virtually every subdiscipline of consciousness science. There are promising application of mathematical concepts in proposals as diverse as quality spaces (Clark, 1993; Rosenthal, 2015; Lee, 2021), qualia spaces (Stanley, 1999), experience spaces (Kleiner & Hoel, 2021; Kleiner & Tull, 2021), qualia structure (Kawakita, Zeleznikow-Johnston, Tsuchiya, & Oizumi, 2023; Kawakita, Zeleznikow-Johnston, Takeda, Tsuchiya, & Oizumi, 2023; Tsuchiya et al., 2022)—cf. Chapter 5.04 of this volume—, Q-spaces (Chalmers & McQueen, 2022; Lyre, 2022), Q-structure (Lyre, 2022), Φ-structures (Tononi, 2015), perceptual spaces (Zaidi et al., 2013), phenomenal spaces (Fink, Kob, & Lyre, 2021), spaces of subjective experience (Tallon-Baudry, 2022), and spaces of states of conscious experiences (Kleiner, 2020a). We will refer to these proposals jointly as proposals of *mathematical structures of conscious experience* in what follows.

Over the previous years, consciousness science has seen a steep increase in the use of mathematical structures of conscious experience. These developments might constitute early signs of a *structural turn* in consciousness science (Kleiner, 2024c), in which mathematical spaces and structures are used, in conjunction with the tools that are available already, to improve theories, experiments, measures and concepts. In what follows, we explain some of the more palpable promises of a structural turn. Needless to say, the review here is very limited. The promises of a structural turn can only be realised by community efforts, spanning researchers across different fields and research programmes.

3.1.1 Structural Theories

A first big impact that structural approaches can have on consciousness science concerns theories of consciousness, specifically the use of mathematical structures or spaces to represent conscious experience and phenomenal character in theories of consciousness, as mentioned already in Section 1.1.

Such 'structural theories' are not different in kind from binary theories of consciousness. Structural approaches do not constrain the metaphysical or conceptual content of theories. They too are hypotheses about how consciousness and the subject matter of the natural sciences relate. The difference between structural and non-structural theories is that the former employ a different way to handle, describe or represent conscious experience, based on mathematical spaces and mathematical structure.

We have already touched on the advantages that structural theories might bring to consciousness science in Section 1.1.3: structural theories are *more explanatory*; and structural theories are *more predictive*.

Further work is necessary to realise these advantages. At the present stage of development, only three structural theories are available: Integrated Information Theory (Chapter 2.07), Higher Order Thought Theories (Chapter 2.03), and Expected Float Entropy Theory (Mason, 2013). But while all of these theories employ or account for *some* structure to represent phenomenal character, they do not yet account for the *actual* structure as found in psychophysical approaches and mathematical phenomenology. Only when they do this will some of the advantages start to realise.

3.1.2 Structural Experiments

A second huge promise of a structural turn in consciousness science concerns experiments. Making use of structure to describe or represent conscious experience is likely to change measurement in consciousness science.

A case in point is the measurement of Neural Correlates of Consciousness (NCC). As shown by (Fink et al., 2021), if certain assumptions about structure hold true, most notably (a) structuralism—the idea that the structure of conscious experiences fully determine all non-structural experiential facts, including which phenomenal properties are instantiated in an experience—, and (b) that there is an isomorphism between the physical structure of the brain and the phenomenal structure of experience, a whole new paradigm to search for the NCCs can be provided. This paradigm might not rely on reports in near-threshold contrast conditions, which are essential for the contrastive analysis approach to NCCs (Baars, 1986).

Unfortunately, as shown in (Kleiner, 2024c), the assumptions presumed in (Fink et al., 2021) are not justified as general conditions on which an NCC research programme can rely. Most notably, the assumption of an isomorphism between the physical and phenomenal domains, or of a structure-preserving mapping more generally, does not serve the purpose it is required to serve in this context. Therefore, the research programme outlined in (Fink et al., 2021) can only be understood as a research programme that presumes a specific class of structural theories of consciousness; this class is to all structural theories of consciousness what the class of identity theories is to all non-structural theories of consciousness.

In spite of these constraints, the general point that Fink et al. (2021) are in essence making—that structural approaches might afford a whole new class of measurement schemes for NCCs, and perhaps also for other measurements in consciousness sci-

ence, one might add (cf. Section 2.3), holds true. The identification of technical difficulties of this first proposals is a sign of good progress, and might lead to ways of overcoming them. The future of NCC research, and measurement of consciousness in general, might well lie in methodologies that combine structural tools with novel experimental or philosophical ideas.

3.1.3 Structural Concepts

The most exciting promise of a structural turn in consciousness science, which (of course) is also the one which is most difficult to assess at the present time, might arguably be the creation of new methodologies for consciousness science, where the term 'methodology' is used in the general sense of a "body of methods used in a particular field of study or activity" (Oxford English Dictionary, 1989). This includes theories of consciousness and novel experimental tools, as discussed in the previous two sections, but may also go beyond them.

Structural approaches may offer entirely new avenues for conceptual engineering. This could be the case, for example, in the context of mathematical phenomenology (cf. Section 3.2), where structural approaches could afford entirely new ways of representing, describing and thinking about aspects of phenomenology. Perhaps mathematical structural approaches can address those aspects of phenomenal experience that are difficult to express in common language, for example nondual awareness or nonegoic reflexivity (Metzinger, 2024), to name just two. And it could be the case in psychophysics, where structural approaches could afford entirely new ways of representing and measuring structural phenomenal properties, new ways which are grounded in the mathematical structure of said properties.

These opportunities are particularly interesting from an illusionist or discourse eliminativist perspective (Frankish & Sklutová, 2022; Irvine & Sprevak, 2020), both of which hold that existing concepts that address the target phenomenon of consciousness are misleading, and should either be discarded or regarded as illusory. However, they do not aim to discard the field of consciousness science entirely, but rather propose alternative concepts. "The positive part of a discourse eliminativist's argument aims to show that an alternative way of talking, thinking, and acting is available" (Irvine & Sprevak, 2020).

Structural concepts might offer such an alternative way of talking and thinking. This is the case because structural concepts can be grounded directly in empirical data, much like is the case in the monumental foundational measurement theory of Krantz et al. (1971). Because of this, structural methodologies could provide a foundation to develop concepts and methods that overcome what appears to be—from the perspective of these views—ill-founded conceptual foundations.

This brief description has only mentioned a small fraction of structural research. There is much more, and it is likely fair to say that the general upshot is that structural approaches have a huge potential in consciousness science. They carry a promise to

vastly extend the range and scope of consciousness science, and might offer a new perspective on many questions currently studied in the field.

In order to realise these promises, structural research needs a thorough foundation. Much work is being directed at establishing such foundations at the moment. At the present time, there is a division of labour, where philosophers are concerned with the analysis of concepts that structural claims should refer to, such as of qualities, cf. (Lee, 2021) and references therein, and where mathematicians are concerned with the analysis of what structural claims *mean*, when applied to such concepts. This is the question of how we should define or understand claims that attribute mathematical structure to qualities or phenomenal character more generally, cf. (Kleiner & Ludwig, 2024) and (Kleiner, 2024b).

3.2 Mathematical Phenomenology & Computational Phenomenology

Particularly noteworthy examples of research on formal concepts and formal methodologies in consciousness science is research on mathematical and computational phenomenology.

The term 'phenomenology' denotes various concepts in consciousness science, cf. Chapter 0.01 of this volume. It is sometimes used to denote the object of investigation in consciousness science, viz. what is also denoted by terms like 'conscious experiences' or referred to by locutions like "what it is like" (Farrell, 1950; Nagel, 1974). But it also denotes a way of engaging with consciousness scientifically and philosophically. Phenomenology in this latter sense refers to a discipline (and movement) in philosophy that contributes to consciousness science, but also has goals that transcend it. It is often taken to be grounded in the works of Edmund Husserl.

Phenomenology has important insights, and important methodologies to offer to consciousness science. In contrast to conceptions of consciousness that are prominent in other disciplines and other parts of philosophy, phenomenology emphasises the lived character of experience, how experience constitutes itself, and structural aspects of a more dynamical nature. All of these are part of the object of investigation of consciousness science, and a full scientific understanding of consciousness will have to include these aspects as well.

3.2.1 Mathematical Phenomenology

Mathematical phenomenology, also called mathematized phenomenology, aims to apply mathematical concepts and techniques in phenomenological investigations, most notably in the form of mathematical presentations of the results of such investigations. Mathematical phenomenology was pioneered by (Petitot, 1999; Casati, 1999) and Yoshimi (2007). Current research in mathematical consciousness science contributes both to the application of mathematical phenomenology, and to the investigation and improvement of its methodology.

Important recent applications of mathematical phenomenology are carried out in Prentner (2019, 2024b), who uses mathematical tools related to pre-topologies to provide a mereological account of the unity of consciousness, intentionality, the self-world distinction, and time.

Work on the foundations of mathematical phenomenology is concerned with the analysis and improvement of its methodology. The investigation of the general conditions of using mathematics to represent or describe conscious experiences in (Kleiner & Ludwig, 2024) is an example. It offers a definition and methodology for how to use mathematics to represent or describe conscious experience based on variations of conscious experiences. Variations can be induced by variations of stimuli that are presented to a subject, as in the case of psychophysical spaces, but they can also result from eidetic variations in Husserl's sense (Husserl, 1939; De Santis, 2011). Eidetic variations are an important method in Phenomenology, and by grounding mathematical representations on this method, the scope of phenomenological investigation can be extended naturally to include mathematical symbolism.

Furthermore, because the definition and methodology can be applied to both variations of stimuli and eidetic variations, the research presented in (Kleiner & Ludwig, 2024) offers the possibility of connecting psychophysics with phenomenology. It offers the hope of unifying quality spaces as constructed "in the lab" in psychophysical measurements with mathematical representations of phenomenology as constructed in phenomenological studies. This could lead to a more comprehensive and thorough representation of conscious experiences in terms of mathematical structure, and offers a hope of cross-inspiration of the two fields. Whether or not these hopes realise in practise is a matter of future study. But it is inspiring to think that mathematics might be the key to arrive at a more unified and integrated agenda in consciousness science, which transcends the boundaries and mutual criticisms of methodologies that exist at the moment.

3.2.2 Computational Phenomenology

A particularly interesting development in the context of mathematical approaches to the mind is computational phenomenology. Computational phenomenology, in its original conception, is the modelling of phenomenology in terms of computational tools from computer science (Harlan, 1984; Petitot, 1999). More recently, the term has been used to denote the modelling of phenomenology in terms of the computational framework provided by Predictive Processing Theory and its Active Inference doctrine (Ramstead et al., 2022), cf. also (Metzinger, 2024).

Because models of computation are mathematical models, computational phenomenology can be seen as a part of mathematical phenomenology. And much like mathematical phenomenology in general, computational phenomenology in particular requires a solid foundation of how to represent phenomenology in mathematical terms. Foundational work on what it means to represent phenomenology mathematically is also foundational work of what it means to represent phenomenology computationally.

In light of this connection between mathematical phenomenology and computational phenomenology, computational phenomenology presents an interesting future application of some of the work on mathematical structures of conscious experience in mathematical consciousness science.

Furthermore, if less standard mathematical structures were found in mathematical phenomenology, it might be the case that research on mathematical structure of Predictive Processing Theory (PP) and Active Inference might become relevant for Computational Phenomenology. The study of this structure in mathematical consciousness science is already underway, as exemplified by (Rudrauf et al., 2023; Sergeant-Perthuis et al., 2023) and (Tull et al., 2023), cf. Section 1.2.2.

3.3 No-Go Theorems in Consciousness Science

An important methodological tool in physics are so-called *no-go theorems*. These are theorems, in the mathematical sense of the term, that establish a conclusion about a subject matter of interest based on mathematical assumptions and a proof. The conclusion usually establishes that something is impossible, hence the 'no-go' in the name. A well-known example in physics is John Bell's proof that local realism—roughly, in this context, the idea that a world composed of localised elements with definite states—cannot be true if quantum theory is true (Bell, 1964).

The idea that no-go theorems might be useful in consciousness science goes back to Ryota Kanai. Inspired by this idea, the research in (Kleiner & Ludwig, 2023), (Kleiner & Hoel, 2021), and (Kleiner, 2024a) made use of the methodology of no-go theorems.

Fundamentally, the idea behind research based on no-go theorems in consciousness science is the very same as in physics: to make use of assumptions that are mostly uncontested, so as to derive, with mathematical rigour, a proof of a statement which "is a bombshell—hardly anyone would have guessed" (Edgington in response to (D. Lewis, 1976), quoted in (Leitgeb, 2013)), ideally speaking.

In the case of (Kleiner & Ludwig, 2023), the assumptions concern general details related to the design and manufacturing of the chips on which contemporary AI systems run (CPUs, GPUs, etc.), as well as the assumption that consciousness is dynamically relevant. The unintuitive result that the no-go theorem establishes is that contemporary AI systems cannot be conscious. In (Kleiner & Hoel, 2021), to give another example, the assumptions comprise the general mathematical form of contemporary neuroscientific theories of consciousness, as well as the relation between theories of consciousness and measures of consciousness. The unintuitive result that follows is that theories of consciousness cannot be falsified, cf. Section 1.1.2.

Formal theorems are of course well-known in philosophy under the banner of mathematical philosophy, cf. Section 3.4, and the application of formal methods and formal theorems carries as much potential for consciousness science as it does for philosophy. Most notably,

> "it forces us to put our cards on the table, that is, to make tacit presuppositions explicit; it helps us to separate the essential from the accidental by making transparent what exactly is needed to make an argument go through; where two areas (...) share enough mathematical structure, it may allow us to translate arguments in the one area into arguments in the other; it functions as a means by which we can put some of our "intuitions" to the test and correct our epistemic biases (...); it facilitates the illustration of abstract circumstances by means of diagram (...), it forges unexpected connections from philosophy to those scientific areas in which mathematical methods are accepted as a standard anyway" (Leitgeb, 2013, p. 274).

In the case of no-go theorems, in addition to establishing a particular conclusion, no-go theorems also serve a second purpose, a purpose which is somewhat implicit in physics but which should be made explicit in consciousness science: they shift attention and resources from the subject matter addressed in the conclusion of the theorem to the subject matter addressed by the assumptions. Correspondingly, a major conclusion of (Kleiner & Ludwig, 2023), for example, is that more attention needs to be placed on studying the substrate on which AI systems run. A major conclusion of (Kleiner, 2024a) is that more research is needed to understand the novel concept of mortal computation that is emerging at the present time; and the conclusion of (Kleiner & Hoel, 2021) is that the present paradigm of formulating and testing theories of consciousness needs to be modified, cf. Sections 1.1.2. The hope in using no-go theorems in consciousness science, therefore, is also to shift attention, and potentially resources, to the assumptions that feed into a theorem.

Mathematical methods and formal theorems have been very useful in philosophy, and no-go theorems have made a large impact to the development in physics. It is likely that they can also play a noteworthy role in making progress in consciousness science, and it would be nice to see further explorations of this opportunity.

3.4 Mathematical Philosophy of Mind

Much of the research on formal concepts and formal methodologies in consciousness science, and much of the research on mathematical consciousness science in general, falls under the banner of mathematical philosophy of mind. *Mathematical philosophy* is "the application of mathematical methods to philosophical questions and problems" (Leitgeb, 2013, p. 269). *Mathematical philosophy of mind*, correspondingly, is the application of mathematical methods to philosophical questions and problems of the mind.

Following Leitgeb (2013)'s explication of the role, tasks and opportunities of scientific philosophy, of which mathematical philosophy is a part, mathematical philosophy of mind can be understood in three equally valid ways. Firstly, it can be understood as *philosophy for the mind sciences*. That is, a philosophy that reflects on the developments in the mind sciences which are mathematical in nature, on a meta-level, with the goal of reforming or improving those developments. Secondly, it can be understood as *philosophy that is part of the mind sciences*. That is, a philosophy which works hand-in-hand with scientists were mathematical methods appear, are applied, or could be helpful in the mind sciences. According to this

understanding, mathematical philosophy of mind works with the same object languages as the sciences of the mind, but addresses some of the more general and more fundamental mathematics-related questions that appear. Thirdly, mathematical philosophy of mind can be understood as *philosophy of mind done with mathematical methods*. According to this understanding, mathematical philosophy of mind targets the questions, problems and hypotheses that are unique to philosophy of mind, but uses formal and mathematical methods to do so, for example formal explications of non-formal concepts. Mathematical philosophy of mind, according to this last understanding, does not aim to improve progress in the mind sciences, but rather is primarily concerned with the long-standing questions of philosophy of mind.

Philosophical research in mathematical consciousness science mostly falls into the second category, it is mathematical philosophy of mind carried out as part of the mind sciences. Inspired by Metzinger (2007)'s analysis of the types of interaction between philosophy of mind and the mind sciences, the intersection between mathematical consciousness science and mathematical philosophy of mind can be classified as follows.

Analysis of the target phenomenon. An important role of philosophy of mind in relation to the sciences of the mind, and correspondingly of mathematical philosophy of mind, can be subsumed under the header of analysis of the target phenomenon. This includes, mainly, the study of concepts that refer to the target phenomenon, so as to make the phenomenon, or properties thereof, accessible to scientific investigations. Such analysis can include both work on existing concepts and the proposal of new concepts, and aims to improve progress in the sciences either by making these concepts available for use in scientific investigations, or by helping scientists to conceptualise the problem under consideration.

Pivotal examples of such analyses in consciousness science are the introduction and analysis of qualia (C. I. Lewis, 1929; Peirce, 1866; Dennett, 1988; Shoemaker, 1991; Block, 2004), of phenomenal consciousness (Chalmers, 1996; Husserl, 1960/2013), of the 'what it is like to be' characterisation of experience (Farrell, 1950; Nagel, 1974), and of the hard problem of consciousness (Chalmers, 1995) and the notion of an explanatory gap (Levine, 1983). The last two have played a particularly important role in shaping what scientists think about their object of investigation, independently of whether they are in agreement with these analyses or not.

Mathematical philosophy of mind is particularly suitable to carry out such analyses in cases where formal concepts are applied to the target phenomenon. In the context of consciousness, this is the case, for example, for mathematical representations of conscious experience or phenomenal character, as discussed in Section 3.1.3.

Analysis of methods. A second pillar of how philosophy of mind supports the mind sciences is the analysis of methods used by the latter, including critical analysis of existing methods, so as to investigate whether these methods work as intended, constructive analyses of existing methods, so as to improve these methods, and the proposal of new methods. Examples in consciousness science are the critical analysis of measures of consciousness in (Irvine, 2012) or (Michel, 2019) and the constructive analysis of C-Tests in light of natural kinds in (Bayne et al., 2024).

Mathematical philosophy of mind offers two avenues of extending this work. On the one hand, it can cope with methods in the sciences which are thoroughly mathematical in nature. On the other hand, it can apply mathematics to provide new analyses of methods of the sciences which are not mathematical in nature. Examples of such analyses are the analysis of new methods to search for NCCs mentioned briefly in Section 3.1.2, or, regarding the applications of mathematics to analyse methods which are not mathematical in nature, the analysis of the contemporary paradigm for testing theories of consciousness discussed in Section 1.1.2.

Analysis of results. A third pillar of how philosophy of mind interacts with the mind sciences is the analysis of results of investigations of the mind sciences. This includes results of empirical studies as well as results of theoretical investigations, for example of theories of consciousness. The goal is to improve the understanding of such results, for example by critical analysis of whether the conclusions that are being drawn are justified, or by constructive analyses of what the result might mean or imply. Examples of this mode of interaction in the case of consciousness science are the numerous analyses and criticisms of theories of consciousness, and the analyses of implications of experiments for the larger questions of the field.

Mathematical philosophy of mind can extend this interaction to results which are thoroughly mathematical in nature. The main examples thereof, to date, are analyses and reconstructions of the mathematical structure of theories of consciousness, as reviewed in Section 1.2.

Bottom-up constraints. A fourth mode of interaction between philosophy of mind and the sciences of the mind, which Metzinger (2007) identifies (cf. also (Metzinger, 2024), concerns the use of scientific results as bottom-up constraints on philosophy of mind, so as to inform philosophical research on philosophical questions, for example metaphysical theories. Prime examples in consciousness science are the many philosophical studies of neurological disorders of consciousness, such as blindsight (Cowey & Stoerig, 1991; Stoerig, 2006), agnosia (Devinsky, Farah, & Barr, 2008), dissociative identity disorders (Kihlstrom, 2005) or neglect (Bisiach, Luzzatti, & Perani, 1979).

This mode of interaction is more aligned with the third way of understanding mathematical philosophy of mind mentioned above, and there are no obvious examples at the present stage of development. Perhaps the formal exposition of structuralist assumptions in (Kleiner, 2024c), and the assessment of such assumptions in light of the types of structures that appear in the sciences could be taken to be an example, but knowledge about the mathematical structures of conscious experience is too limited at the present point of development to draw any thorough metaphysical conclusions.

The exploration of mathematical philosophy of mind, however understood, is likely a very worth-while enterprise. “The desiderata of exactness and fruitfulness will always ‘pull’ explication towards the application of mathematical methods” (Leitgeb, 2013, p. 272). The big promise of mathematical philosophy of mind to consciousness science is that it can help facilitate the important interaction between

philosophy and consciousness science also in the novel phase of mathematized consciousness research.

4 Research on Artificial Consciousness

In light of the vast developments of Artificial Intelligence (AI) in recent years, questions pertaining to a mind of artificial systems have become particularly important. Due to its ethical (Metzinger, 2021), legal and societal relevance, the question of whether artificial systems are or can be conscious, referred to as the question of *synthetic phenomenology* or *artificial consciousness*, is in need of particular attention.

Because AI systems are mathematical systems—they are defined by formal or mathematical structures, both on the level of programming and the level of machine code—the question of synthetic phenomenology is particularly amenable to mathematical tools. To apply a theory or concept to an AI system, the theory or concept needs to be flashed out in formal details. Hence, artificial consciousness has become a major topic of interest in mathematical consciousness science.

4.1 Research Questions

Investigations of the potential of Artificial Intelligence (AI) to exhibit conscious experiences, and of the nature of those experiences where it is indeed possible, are starting to become a key area of investigation in consciousness science. The list of questions that this area of investigation will have to answer is long. It comprises, for example, the following questions:[8]

1. **Are AI systems conscious?** – Or to be more precise, which AI systems are conscious? This is the question of whether AI systems can have conscious experiences at all, independently of what the experiences are.
2. **Are there tests for whether AI systems are conscious?** – This is the question of whether there are operational procedures that can be applied to AI systems so as to infer whether they are capable of having conscious experiences. As of late, such procedures have come to be called *C-tests* (Bayne et al., 2024). Simple examples, like direct interpretation of what Large Language Models (LLMs) state about their own experiences, are not suitable for rigorous tests, simply because LLMs are trained on huge amounts of data that include analyses of and statements about consciousness, so that a suitable prompt, given the appropriate fine-tuning, can lead to corresponding reports independently of whether LLMs are conscious or not. Other tests will have to be found.

[8] I would like to cordially thank Lenore Blum and Ryota Kanai for discussions on the topic. The list of questions presented here came up in a discussion with them. It was also presented at the ASSC27 *Blueprints for Machine Consciousness* symposium in Tokyo.

3. **Are there theoretical means to assess whether AI systems are conscious?** – Here, 'theoretical' includes both scientific and philosophical methods. A particularly important question in this class is:
4. **Can scientific theories of consciousness provide reliable assessments of consciousness in AI systems?** – The emphasis here is on 'reliable'. Theories of consciousness are hypotheses about how conscious experience and the subject matter of the sciences relate. The question is whether a theory of consciousness that targets the subject matter of the brain sciences can also be applied rigorously to AI systems. Are the theoretical constructs, and the empirical evidence, rigorous and detailed enough to warrant the application of theories of consciousness to AI systems?
5. **Which conscious experiences do AI systems have, when conscious?** – This is the question of the phenomenal character of the conscious experiences of AI systems—the question of what it is like to be an AI system in a particular state. Are the conscious experiences of AI systems anything like human conscious experiences? If so, what are the similarities and differences? If not, is there anything that can be said about AI's experiences? This includes, in particular, the following question:
6. **Can AI systems feel pain?** – Or do they otherwise suffer? This is an important ethical question, without resolution of which there is a potential for humankind to create a tremendous amount of suffering, cf. (Metzinger, 2021).
7. **Artificial Phenomenology** – Is it possible to apply the basics of Phenomenology to artificial systems, so as to develop an understanding of the phenomenology of artificial systems, if such phenomenology exists? Perhaps by use of computational or mathematical phenomenology (Section 3.2), or objective phenomenology (Nagel, 1974; Lee, 2021)?
8. **Are there measures of artificial consciousness?** – This is the question of whether it is possible to construct measures of consciousness that can be applied to AI systems, for example to find out whether AI systems experience a particular stimulus consciously, or not.
9. **Can one build conscious AI systems?** – Do the theories, or tests, provide enough details to create blueprints for building AI systems? Can we implement the precise properties that theories of consciousness pin down as sufficient for consciousness? And if so, do they provide enough details to warrant assessments of the type of conscious experience the AI system will have? Can it be ascertained, for example, that the AI system will not be in pain or constantly suffer, as required by (Metzinger, 2021)?
10. **Which forms of computation are suitable to support artificial consciousness, if any?** – This question comprises two questions: the question of which types of computation can support consciousness—can a Turing computation, say? And the question of which specific properties a computation would have to exhibit so as to support artificial consciousness, if any does.
11. **Which role did evolution play in the emergence of conscious systems?** – And which implications does this have for consciousness of artificial systems?

12. **Does consciousness matter for existential risk?** – This is the question of whether conscious AI systems, in particular self-conscious AI systems, pose a particular worry in the context of alignment and existential threats.

Research in mathematical consciousness science is beginning to contribute to many of these questions. For an assessment of question **1.** based on a model of consciousness from computer science, see Chapter 5.02 of this volume. In what follows, we will provide two examples that, in addition to the research presented in Chapter 5.02, illustrate particularly formal approaches.

4.2 Implications of CPU and GPU Design

An example of mathematical contributions to question **3.** is Kleiner and Ludwig (2023). Contemporary AI systems, for example Generative Pretrained Transformers (GPTs), which include Large Language Models (LLMs), are computer programmes. They consist of a few hundred lines of code (almost nothing compared to the tens of millions of lines of code of operating systems like Windows, macOS, or Linux) and a large file of several hundred gigabytes which only contains numbers.[9] What brings these two things together are processing units (PUs). The numbers are converted to strings of zeros and ones, and the code file, once compiled and executed, instructs the processing unit what to do with these strings. If one runs an AI on one's own computer, the task is done by one's central processing unit (CPU), but more advanced systems usually make use of graphics processing units (GPUs), tensor processing units (TPUs), or, as of late, data processing units (DPUs). All of them perform calculations with the numbers as specified in the code, but they differ in how optimised and effective they are in doing this. Because of this, as far as the physical substrate is concerned, contemporary AI systems actually are processing units (PUs). PUs are what supports an AI, much like brains are what supports you.

In light of this it is very surprising that the nature of PUs have not received any attention in investigations of whether AIs have minds, including questions of AI consciousness, prior to the work carried out in (Kleiner & Ludwig, 2023). That is the case even though PUs are fundamentally different from brains and biological substrate.

The largest difference between biological systems like brains and PUs is that PUs are designed and verified to behave exactly as specified by a formal system in the sense of mathematics: the calculation in terms of zeros and ones on the chip is precisely governed by pre-set mathematical rules. "Artificial", in this case not only means man-made, but it means that the system is made to behave in an exact pre-specified way.

[9] The numbers are the weights of an artificial neural network, which result from a training task. Training is the difficult and expensive part of creating an LLM. Running the LLM is comparably cheap and can, with enough patience, also be done on a personal computer.

The analysis in (Kleiner & Ludwig, 2023) shows that this fact has strong consequences if consciousness is dynamically relevant. Intuitively speaking, this is the case because dynamical relevance requires consciousness to be able to make some difference in its substrate, e.g. in the case of a system's report about its conscious experience. But the exact adherence of a PU to a pre-set formal system ensures this can't happen.

4.3 Mortal Computation

An example of mathematical contributions to questions **10.** and **11.** is (Kleiner, 2024a). It also focuses on the distinction between biological and artificial systems, though not on the level of substrate, as in the research reviewed in the previous section, but rather on the level of computation.

The idea—or better: the observation—that there is a difference between the computation that computers implement, and the computations that biological systems like brains implement, called *neural computation* (Piccinini, 2020), is not new. There are various differences between PUs and brains, and these differences are reflected in models of computation that these systems can instantiate.

There is, however, a deeper difference that goes beyond questions of implementation. This difference was first observed by Hinton (2022), and is called *mortal computation*.

In a nutshell, a computation is mortal if it cannot be separated from the hardware on which it runs; if "it dies with the hardware" (Hinton, 2022, p. 13). All computations carried out by computers to date are immortal, they can be separated from the hardware. In contrast, computations carried out by biological systems are mortal, they cannot be separated from the hardware, because biological computation, which is learned rather than programmed, relies on "large and unknown variations in the connectivity and non-linearities of different instances of hardware" (ibid.). Even if it were possible to copy a mortal computation to another system, it would cease to work.

(Kleiner, 2024a) is a first indication that consciousness may require mortal computation. The chapter shows that computational functionalism—the very idea that consciousness is a computation—implies that consciousness is a mortal computation. That is surprising because the 'computations' in computational functionalism are often conceived of as being Turing computations, examples of which are the programs we run on today's computers and mobile devices. Therefore, the result runs counter to many intuitions. But it is aligned with the original definition of computational functionalism by Putnam (1967), which makes use of probabilistic automata descriptions rather than Turing machines, and considers biological organisms as examples. The result is also surprising because it shows that if computational functionalism were indeed true, then contemporary and near-future AI systems, which are immortal computations, could not be conscious, contrary to thinking in several contemporary analyses, like (Butlin et al., 2023).

It should be emphasised, though, that neither of the results presented in (Kleiner & Ludwig, 2023) or (Kleiner, 2024a) attempt to provide a final answer to the question of artificial consciousness. That is the case because they both only target contemporary and near future systems: contemporary and near-future processing units in the case of (Kleiner & Ludwig, 2023), and contemporary and near-future forms of computation in the case of (Kleiner, 2024a). New developments in the semiconductor industry, for example regarding analogue computations, indicate a trend towards transcending both.

5 Conclusion

No natural science has, so far, been solved without mathematics. And consciousness is a natural phenomenon. Hence it is no surprise that as consciousness science starts to blossom, mathematical questions, problems and tasks come to surface as well.

The goal of this chapter was to introduce the reader to some of the questions, problems and tasks in consciousness science that are amenable to mathematical investigations, combining a review of a (very) small part of the developments in mathematical consciousness science with an attempt to provide an outlook onto the future of the field. The hope is that this introduction showcases that mathematical methods and mathematical tools can be helpful to consciousness science in the theoretical (Section 1), experimental (Section 2), conceptual (Section 3), and methodological (also Section 3) domains, as well as in the investigation of artificial consciousness (Section 4).

It is needless to say, hopefully, that if this goal has been met, it has been met in a hopelessly bias way. There is a host of valuable work in mathematical consciousness science that this introduction hasn't even begun to mention. Engagement with this work is wholeheartedly encouraged.

References

Atmanspacher, H., & Rickles, D. (2022). *Dual-aspect monism and the deep structure of meaning*. Routledge.

Baars, B. J. (1986). What is a theory of consciousness a theory of?—the search for criterial constraints on theory. *Imagination, Cognition and Personality*, *6*(1), 3–23.

Barrett, A. B. (2014). An integration of integrated information theory with fundamental physics. *Frontiers in Psychology*, *5*, 63. Retrieved from `https://www.frontiersin.org/article/10.3389/fpsyg.2014.00063` doi: 10.3389/fpsyg.2014.00063

Barrett, A. B., & Mediano, P. A. (2019). The Phi measure of Integrated Information is not well-defined for general physical systems. *Journal of Consciousness Studies*, *26*(1-2), 11–20.

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., ... others (2024). Tests for consciousness in humans and beyond. *Trends in cognitive sciences*.

Beebee, H., Hitchcock, C., Menzies, P. C., & Menzies, P. (2009). *The Oxford handbook of causation*. Oxford Handbooks Online.

Bell, J. (1964). On the einstein podolsky rosen paradox. *Physics Physique Fizika*, *1*(3), 195.

Benjamin, S. F., & Kob, L. (2023). Can structuralist theories be general theories of consciousness? In *Conscious and unconscious mentality* (pp. 112–129). Routledge.

Bisiach, E., Luzzatti, C., & Perani, D. (1979). Unilateral neglect, representational schema and consciousness. *Brain*, *102*(3), 609–618.

Block, N. (2004). Qualia. In R. L. Gregory (Ed.), *Oxford companion to the mind.* Oxford: Oxford University Press.

Blum, L., & Blum, M. (2022). A theory of consciousness from a theoretical computer science perspective: Insights from the conscious turing machine. *Proceedings of the National Academy of Sciences*, *119*(21), e2115934119.

Blum, M., & Blum, L. (2021). A theoretical computer science perspective on consciousness. *Journal of Artificial Intelligence and Consciousness*, *8*(01),

1–42.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

Buckley, C. L., Kim, C. S., McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: A mathematical review. *Journal of Mathematical Psychology*, *81*, 55–79.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... others (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*.

Casati, R. (1999). Formal structures in the phenomenology of motion. In J. Petitot, F. J. Varela, B. Pachoud, & J.-M. Roy (Eds.), *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science.* Stanford: Stanford University Press.

Cerullo, M. A. (2015). The problem with Phi: a critique of Integrated Information Theory. *PLOS Computational Biology*, *11*(9).

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Chalmers, D. J. (1996). *The Conscious Mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Chalmers, D. J., & McQueen, K. J. (2022). Consciousness and the collapse of the wave function. In S. Gao (Ed.), *Consciousness and quantum mechanics.* Oxford University Press.

Clark, A. (1993). *Sensory qualities*. Clarendon Library of Logic and Philosophy.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*(3), 186-190.

Cowey, A., & Stoerig, P. (1991). The neurobiology of blindsight. *Trends in Neurosciences*, *14*(4), 140–145.

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: from neuronal architectures to clinical applications. *Characterizing consciousness: From cognition to the clinic?*, 55–84.

Dennett, D. C. (1988). Quining qualia. In A. Marcel & E. Bisiach (Eds.), *Consciousness in Modern Science.* Oxford: Oxford University Press.

Dennett, D. C. (1995). *Darwin's dangerous idea – evolution and the meanings of life*. Penguin Books.

De Santis, D. (2011). Phenomenological kaleidoscope: Remarks on the husserlian method of eidetic variation. *New Yearbook for Phenomenology and Phenomenological Philosophy*, *11*, 16–41.

Devinsky, O., Farah, M. J., & Barr, W. B. (2008). Visual agnosia. *Handbook of Clinical Neurology*, *88*, 417–427.

De Vries, B., & Friston, K. J. (2017). A factor graph description of deep temporal active inference. *Frontiers in computational neuroscience*, *11*, 95.

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.

Farrell, B. A. (1950). Experience. *Mind*, *59*(234), 170–198.

Ferguson, A., Meyers, C., Bartlett, R., Banister, H., Bartlett, F., Brown, W., ... others (1940). Quantitative estimates of sensory events. report of the british association for the advancement of science. *The Advancement of Science*, *2*, 331–349.

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*.

Frankish, K., & Sklutová, K. (2022). Illusionism and its place in contemporary philosophy of mind. *Human Affairs*, *32*(3), 300–310.

Ganesh, N. (2020). C-wars: the unfolding argument strikes back–a reply to 'falsification & consciousness'. *arXiv preprint arXiv:2006.13664*.

Hanson, J. R., & Walker, S. I. (2019). Integrated information theory and isomorphic feed-forward philosophical zombies. *Entropy*, *21*(11), 1073.

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

Hanson, J. R., & Walker, S. I. (2023). On the non-uniqueness problem in integrated information theory. *Neuroscience of Consciousness*, *2023*(1), niad014.

Harlan, R. M. (1984). Towards a computational phenomenology. *Man and World*, *17*(3), 261–277.

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? towards a principled account of spatial experience. *Entropy*, *21*(12), 1160.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint arXiv:2212.13345*.

Hoffman, D. D., & Prakash, C. (2014). Objects of consciousness. *Frontiers in Psychology*, 577.

Humphrey, N. (2023). *Sentience: The invention of consciousness*. MIT Press.

Husserl, E. (1939). *Experience and judgment*. Northwestern University Press, 1973.

Husserl, E. (1960/2013). *Cartesian meditations: An introduction to phenomenology*. Springer.

Irvine, E. (2012). *Consciousness as a scientific concept: A philosophy of science perspective* (Vol. 5). Springer.

Irvine, E., & Sprevak, M. (2020). Eliminativism about consciousness. In U. Kriegel (Ed.), *Oxford handbook of the philosophy of consciousness.* Oxford: Oxford University Press.

Kanai, R., & Fujisawa, I. (2024, 05). Toward a universal theory of consciousness. *Neuroscience of Consciousness*, *2024*(1), niae022.

Kawakita, G., Zeleznikow-Johnston, A., Takeda, K., Tsuchiya, N., & Oizumi, M. (2023). Is my" red" your" red"?: Unsupervised alignment of qualia structures via optimal transport. *PsyArXiv: h3pqm*.

Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., & Oizumi, M. (2023). Comparing color similarity structures between humans and llms via unsupervised alignment. *arXiv preprint arXiv:2308.04381*.

Kihlstrom, J. F. (2005). Dissociative disorders. *Annu. Rev. Clin. Psychol.*, *1*, 227–253.

Kleiner, J. (2020a). Brain states matter. A reply to the unfolding argument. *Consciousness and Cognition*, *85*.

Kleiner, J. (2020b). Mathematical models of consciousness. *Entropy*, *22*(6).

Kleiner, J. (2024a). Consciousness qua mortal computation. *arXiv preprint arXiv:2403.03925*.

Kleiner, J. (2024b). The newman problem of consciousness science. *PhilArchive Preptint: KLETNP-4*.

Kleiner, J. (2024c). Towards a structural turn in consciousness science. *Consciousness and Cognition*, *119*.

Kleiner, J., & Hartmann, S. (2023). The closure of the physical, consciousness and scientific practice. *arXiv preprint arXiv: 2110.03518*.

Kleiner, J., & Hoel, E. (2021). Falsification and consciousness. *Neuroscience of Consciousness*, *2021*(1).

Kleiner, J., & Ludwig, T. (2023). The case for neurons: A no-go theorem for consciousness on a chip. *Earlier arXiv preprint arXiv:2304.05077*.

Kleiner, J., & Ludwig, T. (2024). What is a mathematical structure of conscious experience? *Synthese*, *203*(3), 89.

Kleiner, J., & Tull, S. (2021). The mathematical structure of Integrated Information Theory. *Frontiers in Applied Mathematics and Statistics*, *6*.

Kob, L. (2023). Exploring the role of structuralist methodology in the neuroscience of consciousness: a defense and analysis. *Neuroscience of Consciousness*, *2023*(1), niad011.

Krantz, D., Luce, D., Suppes, P., & Tversky, A. (1971). *Foundations of measurement, vol. i-iii*. Academic Press.

Kuhn, R. L. (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in Biophysics and Molecular Biology*, *190*, 28-169. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0079610723001128` doi: https://doi.org/10.1016/j.pbiomolbio.2023.12.003

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lee, A. Y. (2021). Modeling mental qualities. *Philosophical Review*, *130*(2), 263–298.

Leitgeb, H. (2013). Scientific philosophy, mathematical philosophy, and all that. *Metaphilosophy*, *44*(3), 267–275.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, *64*(4), 354–361.

Lewis, C. I. (1929). *Mind and the World-Order: Outline of a theory of knowledge*. Courier Corporation.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. In *Ifs: Conditionals, belief, decision, chance and time* (pp. 129–147). Springer.

Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of mathematical psychology*, *1*(1), 1–27.

Lyre, H. (2022). Neurophenomenal Structuralism. A philosophical agenda for a structuralist neuroscience of consciousness. *Neuroscience of Consciousness*, *2022*(1), niac012.

Mason, J. W. (2013). Consciousness and the structuring property of typical data. *Complexity*, *18*(3), 28–37.

Mason, J. W. (2021). Model unity and the unity of consciousness: Developments in expected float entropy minimisation. *Entropy*, *23*(11), 1444.

Metzinger, T. (2007). *Philosophie des Bewusstseins*. Auditorum Netzwerk. DVD.

Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. *Journal of Artificial Intelligence and Consciousness*, *8*(01), 43–66.

Metzinger, T. (2024). *The elephant and the blind: the experience of pure consciousness: philosophy, science, and 500+ experiential reports*. MIT Press.

Michel, M. (2019). The mismeasure of consciousness: A problem of coordination for the perceptual awareness scale. *Philosophy of Science*, *86*(5), 1239–1249.

Michel, M. (2023). Confidence in consciousness research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *14*(2), e1628.

Michel, M., & Lau, H. (2020). On the dangers of conflating strong and weak versions of a theory of consciousness. *Philosophy and the Mind Sciences*, *1*(II).

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept* (Vol. 53). Cambridge University Press.

Miller, M., Clark, A., & Schlicht, T. (2022). Predictive processing and consciousness. *Review of Philosophy and Psychology*, *13*(4), 797–808.

Moon, K., & Pae, H. (2018). Making sense of consciousness as integrated information: Evolution and issues of iit. *arXiv preprint arXiv:1807.02103*.

Myrvold, W. (2022). Philosophical Issues in Quantum Theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2022/entries/qt-issues/`.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 435–450.

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*, 1–18.

Nida-Rümelin, M. (2018). The experience property frame work: a misleading paradigm. *Synthese*, *195*, 3361–3387.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology*, *10*(5), e1003588.

Oxford English Dictionary. (1989). methodology (n.). In *Second edition.* Oxford University Press.

Ozawa, M. (1984). Quantum measuring processes of continuous observables. *Journal of Mathematical Physics*, *25*(1), 79–87.

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press.

Peirce, C. S. (1866). Lowell lecture. In M. H. Fisch (Ed.), *Writings of charles s. peirce: A chronological edition.* Indiana University Press.

Petitot, J. (1999). Morphological eidetics for a phenomenology of perception. In J. Petitot, F. J. Varela, B. Pachoud, & J.-M. Roy (Eds.), *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science* (pp. 330–371). Stanford: Stanford University Press.

Piccinini, G. (2020). *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford University Press. doi: https://doi.org/10.1093/oso/9780198866282.001.0001

Prentner, R. (2019). Consciousness and topologically structured phenomenal spaces. *Consciousness and Cognition*, *70*, 25–38.

Prentner, R. (2024a). Category theory in consciousness science: going beyond the correlational project. *Synthese*, *204*(2), 69.

Prentner, R. (2024b). Mathematized phenomenology and the science of consciousness. *OSF Preprint*.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion.* Pittsburgh: University of Pittsburgh Press. (Reprinted in (Putnam, 1975).)

Putnam, H. (1975). The nature of mental states. In *Mind, language, and reality: Philosophical papers* (Vol. ii). Cambridge: Cambridge University Press.

Ramstead, M. J., Seth, A. K., Hesp, C., Sandved-Smith, L., Mago, J., Lifshitz, M., ... others (2022). From generative models to generative passages: a computational approach to (neuro) phenomenology. *Review of Philosophy and Psychology*, *13*(4), 829–857.

Rosenthal, D. (2010). How to think about mental qualities. *Philosophical Issues*, *20*, 368–393.

Rosenthal, D. (2015). Quality spaces and sensory modalities. In P. Coates & S. Coleman (Eds.), *Phenomenal qualities: sense, perception, and consciousness* (pp. 33–65). Oxford University Press Oxford, UK.

Rudrauf, D., Sergeant-Perthuis, G., Tisserand, Y., Poloudenny, G., Williford, K., & Amorim, M.-A. (2023). The projective consciousness model: projective geometry at the core of consciousness and the integration of perception, imagination, motivation, emotion, social cognition and action. *Brain Sciences*, *13*(10), 1435.

Sergeant-Perthuis, G., Rudrauf, D., Ognibene, D., & Tisserand, Y. (2023). Action of the euclidean versus projective group on an agent's internal space in curiosity driven exploration: A formal analysis. *Artific Intellig*.

Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in cognitive sciences*, *12*(8), 314–321.

Seth, A. K., & Hohwy, J. (2021). Predictive Processing as an empirical theory for consciousness science. *Cognitive Neuroscience*, *12*(2), 89–90.

Shoemaker, S. (1991). Qualia and consciousness. *Mind*, *100*(4), 507–524.

Signorelli, C. M., Szczotka, J., & Prentner, R. (2021). Explanatory profiles of models of consciousness-towards a systematic classification. *Neuroscience of*

consciousness, *2021*(2), niab021.

Signorelli, C. M., Wang, Q., & Khan, I. (2021). A compositional model of consciousness based on consciousness-only. *Entropy*, *23*(3), 308.

Stanley, R. P. (1999). Qualia space. *Journal of Consciousness Studies*, *6*(1), 49–60.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Stoerig, P. (2006). Blindsight, conscious vision, and the role of primary visual cortex. *Progress in Brain Research*, *155*, 217–234.

Tallon-Baudry, C. (2022). The topological space of subjective experience. *Trends in Cognitive Sciences*.

Tononi, G. (2015). Integrated Information Theory. *Scholarpedia*, *10*(1), 4164.

Tsuchiya, N., Andrillon, T., & Haun, A. (2019). A reply to "the unfolding argument": Beyond functionalism/behaviorism and towards a truer science of causal structural theories of consciousness. *PsyArXiv*.

Tsuchiya, N., Phillips, S., & Saigo, H. (2022). Enriched category as a model of qualia structure based on similarity judgements. *Consciousness and Cognition*, *101*, 103319.

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), niab034.

Tsuchiya, N., Saigo, H., & Phillips, S. (2023). An adjunction hypothesis between qualia and reports. *Frontiers in Psychology*, *13*, 1053977.

Tull, S., & Kleiner, J. (2021). Integrated Information in Process Theories: Towards categorical IIT. *Journal of Cognitive Science*, *22*(2), 92–123.

Tull, S., Kleiner, J., & Smithe, T. S. C. (2023). Active inference in string diagrams: A categorical account of predictive processing and free energy. *arXiv preprint arXiv:2308.00861*.

von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik (Mathematical Foundations of Quantum Mechanics)*. Julius Springer.

Yoshimi, J. (2007). Mathematizing phenomenology. *Phenomenology and the Cognitive Sciences*, *6*(3), 271–291.

Zaidi, Q., Victor, J., McDermott, J., Geffen, M., Bensmaia, S., & Cleland, T. A. (2013). Perceptual spaces: mathematical structures to neural mechanisms. *Journal of Neuroscience*, *33*(45), 17597–17602.