

Contents

Research Articles

Stefan Petkov: <i>The Degrees of Understanding and the Inferential Component of Understanding</i>	746
Yavuz Recep Başoğlu: <i>How Not to Argue about the Compatibility of Predictive Processing and 4E Cognition</i>	777
Serdal Tümkaya: <i>A Novel Reading of Thomas Nagel's "Challenge" to Physicalism</i>	802
Vojtěch Zachník: <i>Epistemic Foundations of Salience-Based Coordination</i>	819
José Ángel Gascón: <i>Why Did You Really Do It? Human Reasoning and Reasons for Action</i>	845
Guy Dove – Andreas Elpidorou: <i>A Dilemma about the Mental</i>	867

Miscellanea

Pekka Väyrynen: <i>Erratum: Normative Naturalism on Its Own Terms</i>	896
<i>Contents 2021</i>	897

The Degrees of Understanding and the Inferential Component of Understanding


Stefan Petkov*

Received: 22 August 2019 / Revised: 11 October 2020 / Accepted: 28 September 2020

Abstract: Current debates on the nature of explanatory understanding have converged on the idea that at least one of the core components of understanding is inferential. Philosophers have characterized the inferential dimension of understanding as consisting of several related cognitive abilities to grasp a given explanation and the nexus of complementing explanations to which it belongs. Whilst analyses of both the subjective epistemic abilities related to grasping and objective features of the inferential links within explanations have received much attention, both within theories of explanation and in the literature on understanding, the criteria for evaluating the specific structure and organization of explanatory clusters or nexuses has received much less attention. Nevertheless, two notable exceptions stand out—Khalifa’s characterization of an explanatory nexus, and theories of explanatory unification. I take Khalifa’s ideas, together with the basic criteria of successful explanatory unification, as my starting point. To both I make some corrections and additions, in order to arrive at a more robust notion of an explanatory nexus and ultimately show that its structural properties and the inter-explanatory relations it contains are relevant to the resulting understanding.

* Beijing Normal University

 <https://orcid.org/0000-0001-9152-574X>

 School of Philosophy, Beijing Normal University, Room A801, Front Main building, Beijing Normal University, No. 19 Xijiekouwai Str, Haidian District, Beijing, China

 yaggdrasil@yahoo.com

© The Author. Journal compilation © The Editorial Board, *Organon F*.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0).

I propose to represent such nexuses as directed graph trees and show that some of their properties can be related to the degree of understanding that such nested explanatory structures can offer. I will further illustrate these ideas by a case study on an ecological theory of predation.

Keywords: Ecological theory of predation; explanatory understanding; scientific explanations; unification.

1. Introduction

The problem of explanatory understanding has received significant attention in the recent literature of epistemology and philosophy of science (see Baumberger et al. 2017 for an overview). The exact definition of understanding¹ is a matter of an ongoing dispute, but most analyses have converged on the idea that at least one of the key differences between mere knowledge of a correct explanation of a phenomenon and understanding the phenomenon has an inferential character (Newman 2014; Grimm 2010; Khalifa 2017; De Regt 2015). The inferential character of the explanatory understanding of a given fact, or a factual domain, has been analysed in the literature in two complementary ways, related to both the inferential properties of singular explanations and the organization of inter-linked explanations:

1. As a subjective ability to evaluate the quality of the available explanations (Grimm 2010; Newman 2014; Khalifa 2017), the ability to grasp the explanations (e.g. arriving at new inferences based on the original explanatory blueprint) and to relate these explanations together forming an explanatory nexus;
2. As the objective features of explanatory inferences in terms of the type of relation between the explanans and the explanandum (Kostić

¹ For stylistic purposes, here I will use “understanding” and “explanatory understanding” interchangeably to refer to the same epistemic good of comprehending a phenomenon by an available explanation. A further clarification on this notion can be found in section 2.

2018), the informativeness and relevance of the explanatory premises, and the quality of the so-called cluster or nexus of complementing explanations (Khalifa 2017).

If we take into account the idea that understanding is not a binary notion, but comes in degrees (Khalifa 2017; Newman 2014), we can then see that the degree of understanding of some fact or factual domain depends on both subjectively grasping the conceptual inferential relations within explanations or between complementing explanations of that fact or domain of facts, but also on objective features of these explanations and established inter-explanatory relations. Whilst analyses of the subjective epistemic abilities related to grasping, and the specific features of the inferential links within explanations, have received extensive attention, both in theories of explanation (Hempel 1965; Lewis 2000; Woodward 2003; Lange 2016) and in the literature on understanding (Baumberger et al. 2017), criteria for evaluating the specific structure and organization of explanatory clusters or nexuses has received much less attention.

Two notable exceptions stand out. In the literature on understanding Khalifa has introduced the general notion of an explanatory nexus (Khalifa 2017). He has suggested that the degree of understanding that an explanatory nexus provides is determined by comparative principles of completeness and likeness to scientific knowledge. However, his characterization of explanatory nexuses and the criteria for their evaluation have remained too broadly defined.² In theories of explanations, unificationist theories have also conveyed the idea that explanations should not be considered in isolation. These theories have developed the idea that understanding should be judged both on the basis of the number of facts an explanation can cover (Kitcher 1989) and on the number of other explanations that an explanation can unify (Friedman 1974; Bartenbloth 2002). Therefore, unification can potentially provide the grounds for a more robust criterion for the evaluation of a cluster of explanations, in terms of their organization and their inter-explanatory relations.

² Khalifa's analysis of completeness of grasp for instance is overly broad and as he himself notes is not susceptible to quantitative assessment (see Khalifa 2017, 10), for more details see section 2.3.

Here, I will take both Khalifa's idea of an explanatory nexus and the basic criteria of successful explanatory unification as my starting point. In both cases, I will make some corrections and build upon the existing views in order to define the notion of an explanatory nexus more robustly and to offer more precise criteria for its evaluation. My central effort will be to show that the organizational and structural properties of these nexuses also play an important role in determining the degree of understanding.

To give an outline of the paper. In the second section, I will present a short review of the literature on the subjective and objective features of inferential explanatory understanding. In the third section, I will link the core ideas of these analyses with explanatory unification. In the fourth section, I will lay the groundwork for a possible formal explication of an explanatory nexus as a directed graph tree. Making some corrections to both Khalifa's notion of a nexus and the basic criteria of unification, I will provide robust criteria of the degree of understanding that such nexuses can provide in terms of both the *quantitative evaluation of subjective completeness of grasp* and the *evaluation of the objective properties of explanatory structures*. I will also give a short illustration of these ideas by a case study on an ecological theory of predation.

2.1. The subjective side of the inferential component of understanding

The philosophical analysis of understanding has shown that the difference between mere knowledge of the correct explanation of a given phenomenon and understanding that phenomenon has a distinct subjective dimension. Typically, this dimension has been analysed under the concept of grasping. Grimm (2010) has described grasping as the ability "to work upwards" (so to speak) from a particular explanation towards the relevant fragment of a background theory, by mastering the general conceptual relations, or the dependence conditions, that are exemplified in the explanation.

An extension of the analysis of the subjective epistemic requirements of understanding has been made by Newman (2014). For him, understanding involves several distinct but related cognitive abilities such as semantic

ability, comprehension ability and problem-solving ability. A test for these cognitive abilities would be to check the subject's ability to link the newly introduced explanatory inferences to pre-existing explanatorily relevant information, in order to solve novel problems.

To illustrate, reaching understanding of the phenomenon of flight, by learning the Bernoulli principle, as a result of answering the explanation-seeking question "Why can the plane fly?", is not equivalent to merely knowing that the Bernoulli principle provides relevant information about the conditions which make flight possible. Understanding flight via this explanation means also being able to work out how the variables in the equation will co-vary in different circumstances. "Grasping" this explanation, for instance, will mean being able to infer that the theoretic maximum altitude of a given model of glider depends on the shape of its wings and the density of the air at different altitudes. Understanding in this sense involves the ability to analyse the explanatory relation between the explanans and the explanandum in terms of the conceptual relations exemplified in the explanation and thus arriving at the explanatory blueprint which the explanation exemplifies.

As Newman has shown, grasping will also involve linking the resulting explanatory blueprint to other available explanatory information for the purpose of solving novel problems. For example, linking the newly introduced Bernoulli principle to a background knowledge of thermodynamics will allow us to understand not only flight but also how injectors in steam locomotives work.

Even though our consideration of the idea of grasping clearly shows that explanations should not be considered in isolation, an exclusive focus on the subjective epistemic abilities that make possible the linking of newly introduced explanatory information to pre-existing explanatory knowledge will leave the analysis of understanding one-sided. Obviously, an objective evaluation of both the structural properties of explanatory inferences themselves and of how these inferences are organized together with other background explanatory information (producing larger and more encompassing inferential networks) is also essential.

However, as we shall see, the literature on theories of explanations and on explanatory understanding has focused for the most part only on

analysing the relation between explanations and their explananda and less on providing criteria for the organization of such inferential networks.

2.2. The objective side of the inferential component of understanding and singular explanations

Most³ of the classical literature on scientific explanations has been focused on analysing understanding in terms of the features of the inferential relation between the explanandum and the explanans. This strategy has been endorsed by most well-known theories of scientific explanation, such as the deductive-nomological model (Hempel 1965), and various causal approaches (Dowe 2000; Salmon 1998; Woodward 2003). The deductive-nomological model has been deployed to investigate this relation as a form of logical consecution of the lawful premise in the explanans that subordinates the explanandum; in causal theories of explanation the relation is described as counterfactual dependence (Lewis 2000) or difference making (Woodward 2003) between the cause (explanans) and the effect (explanandum). The introduction of distinctively mathematical explanations has further enriched this picture by also describing some scientific explanations in terms of the relation of logical constraint that a mathematical explanans imposes upon an ontic fact (explanandum) that is satisfied by it (Lange 2016).

Recently, Kostić (2018) has proposed a general criterion for assessing explanatory understanding based on the relation between the explanans and the explanandum. His central claim is that, because the goal of explanation is understanding, understanding is intimately linked with explanation. For him, this intimacy depends on the level of complexity in the structure of explanation in such a way that the more immediate the relation between the explanans and explanandum, the more immediate the understanding.

As a corollary, he also claims that. The explanation will provide more understanding and depth if the explanatory relation is more immediate.⁴

³ With the notable exception on theories of explanatory unification. See section 3.

⁴ Kostić has suggested that the most immediate relation between explanans and its explanandum is non-inferential in character and any type of explanatory inferences such as deductive-nomological explanations are necessarily more complex.

These results from the fact that the immediacy also has as a by-product the possibility to apply that explanation to a large number of phenomena.

He defends these ideas by comparing minimal structure explanations (a type of distinctively mathematical explanations) with deductive-nomological or causal mechanistic explanations. Kostić shows that minimal structure explanations require simply grasping that the explananda satisfy a given mathematical structure and, as such, that the derived properties of that mathematical structure will also be exemplified by the explananda. If these properties are the ones which fall under the spotlight of explanatory seeking problems, then an explanation need only show that the explanandum instantiates the mathematical structure. As a corollary, he suggests that, due to the higher abstractness of these explanations, the explaining mathematical structure can be used to cover a large variety of facts, also offering significant explanatory depth. In contrast to this type of explanation, causal explanations, for instance, have a narrower scope, because they exemplify specific causal interactions and, as such, require specific ontic information which is not generalizable.

Here, I shall agree with the first element of his account—the more intimate the relation between the explanans and the explanandum the more immediate the understanding. The second claim, which links the explanatory depth of minimal structure explanations to the immediacy of the explanatory relation is more contestable and does not fit well with the account of unification I am going to develop here.

Firstly, it is possible that a minimal structure explanation satisfies the norm of intimacy between an explanans and explanandum but remains applicable to a very few phenomena. This is because the number of phenomena that exemplify a given mathematical structure is a contingent fact, which does not depend on the properties of the exemplified mathematical structure. After all, nothing in the immediacy of the relation between a mathematical explanans and a factual explanandum hints that such a relation should be satisfied by more than a single fact. Therefore, it is a contingent

Since the problem I address here is that of inter-explanatory relations and structures, I will leave open the problem of whether there are cases of non-inferential explanations.

ontic issue whether a given mathematical structure, despite its abstractness, is applicable to a large number of facts or not.

Nevertheless, the inverse is certainly correct. The more general the explanation the less specific ontic information it contains and, therefore, the more immediate the relation between its explanans and explanandum. As such, more abstract or general explanations should stand higher in a possible hierarchy of interlinked explanations. As for determining the explanatory depth that such explanations can provide, and use it as a cubit for understanding, I shall not follow Kostić directly. Instead, I will try to develop a more complex view of unification that requires that the general applicability of a given explanation and the explanatory relation it exemplifies should always be taken into account in relation to other more general or more specific explanations.⁵

To conclude, if we take Kostić's account narrowly, as a criterion for evaluating singular explanations by determining the type of relation that holds between the explanans and the explanandum, and by comparing it with different explanatory strategies, it strikes me as fundamentally correct in its basic assumption of intimacy. Therefore, in what follows, I shall take it as a basic criterion for ordering explanations. The more generally applicable the explanation, the more intimacy it will display between its explanans and its possible explananda, and the higher it should stand in a hierarchy of interlinked explanations.

⁵ Broadly applicable explanations (as Kostić himself notes) naturally offer less particular information due to their abstractness. As such they are sometimes supplemented by specific causal mechanistic explanations. The ensuing deeper understanding can then be seen in terms not of broader applicability alone, but as the possibility of linking these more abstract explanations with more narrow causal explanations which offer specific ontic information. Moreover, following Woodward (2003), an account of unification cannot consider that phenomena that are describable by the same mathematical structure are unified e.g. it is possible that the same mathematical equation is used to describe several unrelated phenomena, however we cannot treat these phenomena as well unified under this mathematical equation. For more on the type of unification, I am going to develop see section 3.1.

2.3. Khalifa's combined criterion and explanatory nexuses

In his book, Khalifa (2017) has developed a general analysis of scientific understanding. He has combined both the objective and subjective dimension of understanding by giving the following comparative criterion upon which the degree of understanding of some fact p can be determined:

- (EKS1) $S1$ understands why p better than $S2$ if and only if:
- (A) *Ceteris paribus*, $S1$ grasps p 's explanatory nexus more completely than $S2$; or
 - (B) *Ceteris paribus*, $S1$'s grasp of p 's explanatory nexus bears greater resemblance to scientific knowledge than $S2$'s. (Khalifa 2017, 14)

As we can see from this definition, his criterion relies heavily on the notion of an explanatory nexus, which, for him, is the cluster of correct explanations of p , as well as the relations between those explanations. If we take the second criterion “the resemblance to scientific knowledge” as satisfied—e.g., all the explanations satisfy the accepted objective criteria for the inferential relation between explanans and explanandum—this leaves us with the second element, completeness. However, perhaps due to the general project of his book, Khalifa has left the notion of completeness too broadly defined. Under the heading of completeness, Khalifa has listed both the number of correct explanations and inter-explanatory relations grasped and the quality of these explanations and inter-explanatory relations. Moreover, he has described completeness of nexuses in a non-restricted way—as all the available explanations of a given fact. Even if explanations meet the criteria for scientific likeness (e.g., they are the best our science can offer) these explanations can hardly be organized in any robust way. Theories use many different explanatory strategies, including descriptions of broad mathematical dependencies, deductions derived from lawful generalizations, inferences based on phenomenological modelling; they may include approximations and idealizations, and so on. Without any criteria of the relations holding between these different explanatory products, the notion of a nexus remains (though intuitively appealing) very vague. Consequently, the notion of completeness, as Khalifa defines it, only enables a comparison of the

relative degrees of understanding of different epistemic subjects in simple cases where one of the subjects lacks an explanation that the other possesses (Khalifa 2017, 10).

However, as we shall see, inter-explanatory relations between explanations can be defined more clearly and, given this, both the organization and the completeness of such nexuses can be more precisely described. If this is the case, not only the number of correct explanations, but also their organization and the contrast class of phenomena that they consider, will all be relevant to the resulting understanding.

Khalifa's definition provides a strong starting point. However, in order to render the problem of evaluating explanatory nexuses more manageable, I will narrow my focus to explanatory nexuses limited to a single theory. I will also assume that the examined explanatory structures contain only genuine scientific explanations of one particular type (e.g. that these explanations are all lawful deductions, or describe causal dependences, or are subject to some distinctively mathematical constraints such that all are properly "science-like").

Two problems then stand out. Firstly, *how are we to organize explanatory nexuses*; and secondly, *by which criteria are we to judge the inter-explanatory relations that such nexuses display?*

To make an analogy with the classical theory of explanations: if that theory's central problem has been the characterization of the appropriate types of explanans/explanandum relations, then similarly, for the theory of understanding based on grasping explanatory nexuses, an analysis of explanatory structures grasped in terms of their inter-explanatory relations and organization must also be made.

3.1. Explanatory unification as a starting point for evaluating explanatory nexuses

So far, we have seen that the degree of understanding of some fact or factual domain will depend not only the subjective grasp and objective evaluation of the inferential relations within the available explanations, but also on establishing and evaluating inter-explanatory relations. The problem

then is to present a more precise notion of such structures of explanations, and show how these structures can be evaluated based on their organizational inter-explanatory properties. I believe this idea can be developed on the basis of the existing analyses of explanatory unification. We can start building up a suitable notion of unification by following in chronological order the key accounts of Friedman (1974), Kitcher (1989) and Bartelborth (2002).

The theory of explanatory unification was pioneered by Friedman in 1974. In his original paper, he asked the same question as that raised in the contemporary topic of understanding: in virtue of what do scientific explanations provide understanding? His answer was that scientific theories generate understanding by offering two types of explanations: explanations which reduce the number of independent phenomena by subsuming them under general laws⁶, and more general explanatory inferences that subsume the independent laws. Even though Friedman's account suffered from some technical problems⁷, his central idea of a hierarchy of explanations remains valuable. Roughly, we can paraphrase it as linking the degree of understanding that a cluster of explanations offers to how unified this cluster is. Kitcher (1989), however, did not build directly on this idea of a hierarchy, or interlinked structure of explanations; instead, he focused on describing successful unification in terms of a theory that manages to subsume its domain under the smallest possible set of explanations. To explicate this idea, he developed the concept of an explanatory pattern. Such patterns for Kitcher serve as skeletal blueprints of particular deductive explanatory arguments.⁸ For example, a general pattern for evolutionary dynamics can be:

SNS:

*Explanans: x and y are members of n , differentiated by a variant on an inherited trait z ; x and y are competing in E , where x is fitter than $y \rightarrow$
Explanandum: x 's size is larger than y in E in some $n' > n$).*

⁶ Friedman worked under the tenets of the deductive-nomological model.

⁷ It accepted simple conjunctions of independent laws as explanations of these laws (Kitcher 1976).

⁸ Famously, Kitcher has defended a position of deductive chauvinism—all explanations are ultimately reducible to causal explanations. I shall not follow him in this idea. Instead, I will assume explanatory pluralism.

Here x and y are the demographic sizes of sub-populations of the same species, differentiated by some trait z , E is the relevant environmental factors and n designates the generation number. A simple theory of evolution contains only this pattern plus a set of filling instructions for the dummy letters. When these filling instructions are exhausted all the instances of that pattern can be generated and the domain of the theory can be closed. It is clear that SNS is highly systematic, because it covers every possible evolutionary scenario, excluding only population states resulting from genetic drift. The downside is that such a general pattern offers very little information and thus understanding of actual evolutionary factors. Therefore, systematization or generality alone are insufficient as a measure of understanding. To amend this issue, Kitcher provided another requirement for successful unification. A unified theory for him is such a structure of explanations that manages not only to *systematize* its domain under a small number of explanatory patterns but also one in which these patterns are as *stringent* as possible (e.g. they provide as much particular information as possible).

Jointly satisfying the two requirements of stringency and systematization has been shown to be a serious problem for Kitcher's account, because they clearly pull in different directions (Morrison 2000). Beside the tension between stringency and systematization, there is also the question (left open by Kitcher) of how organizations that display different numbers of patterns and varying degrees of systematization and stringency can be compared (Kitcher 1989, 435).

Barelbroth (2002) has essayed a resolution of these issues by extending Friedman's original notion of a hierarchy of explanations. He makes use of the formal apparatus of structuralism to propose that unification is successful in generating understanding only if it manages to embed the explananda into a hierarchy of explanatory models each of which displays a different level of generality. His and subsequent accounts of unification (Petkov 2015) describe this hierarchy as a treelike structure, at the root of which lie the most general explanatory patterns. These most abstract patterns present the core concepts of the theory, their relations, and how they can be applied in order to map the factual domain (Petkov 2015). These explanatory patterns are then progressively extended by specifying the core conceptual links in such a way that more and more specific explanations can be constructed.

Optimally, this should result in a series of progressively more stringent explanations that should exhaust in a contrastive manner the factual domain and also offer deeper understanding.

Such a picture of theories as hierarchies of linked explanations that exhaust a particular domain is suggestive of the notion of an explanatory nexus. However, we cannot directly deploy the criteria of stringency and systematization in the evaluation and description of explanatory nexuses. This is because, firstly, Khalifa has defined explanatory nexuses as collections of the explanations of singular facts, whilst both Grimm's grasping and unification are geared instead to describing explanations and organizations of explanations in terms of their broader applicability to more than a singular explanandum. Recall that both grasping and unification are evaluated in terms of the possibility of using a particular explanation of some fact as an exemplar of an explanatory blueprint that is then applicable to further phenomena similar to the explanandum. Moreover, as we saw, Khalifa defined explanatory nexuses in a non-restricted way—as all the available scientific explanations of the fact under scrutiny. This permits the inclusion of any number and any type of explanations, and, such clusters clearly does not meet the criteria for unified knowledge.

Another serious problem that this time emerges from the side of unification is the so called “ad explanandum” problem (Halonen and Hintikka 1999; Petkov 2015). Halonen and Hintikka (1999) have criticized explanatory unification because it typically presents the embedding of more specific explanations into more general ones, as a type of inferential derivation. Halonen and Hintikka have suggested that this derivation implies that the explananda are considered as deductively closed or static. Whilst in fact particular explanations often include, specific ad explanandum information. This information is case sensitive and similarly to auxiliary conditions of deductive nomological explanations, is not generalizable nor it can be derived directly from the available explanatory information from the higher level explanations. Similarly, the notion of a general explanatory blueprint that is applied serially (e.g. grasped), cannot actually work, because at each application of that general explanatory pattern we would need to supply additional knowledge specific to the particular explananda, which is then effectively equivalent to discovering new explanations.

Both of these issues present a significant challenge to the project of quantitatively comparing explanatory nexuses in terms of grasping and to the development of a qualitative account of such nexuses in terms of their unification.

3.2. An interrogative account of explanatory nexuses limited to a particular domain

Notwithstanding the above, I believe that, given certain corrections to both our idea of unification and Khalifa's notion of a nexus, we shall see that we can render these concepts much more robust, and also provide some guidelines for a quantitative analysis of the degree of understanding that a nexus provides, in terms of its content and structural properties.

We can start with the curious historiographical observation about the literature of unification. Both its proponents and critics seem to have been preoccupied with Kitcher's notion of an explanatory pattern and the problem of how such patterns hang together. However, in his original essay, Kitcher developed the concept of explanatory patterns, along with the requirements for stringency and systematization, as an extension of van Fraassen's pragmatic approach to explanations. Therefore, in order to meet the challenges of explicating explanatory nexuses more clearly, I will, so to speak, "go back to the roots".

I take Kitcher's basic idea that a systematization of a domain is in fact achieved by patterns which are recoverable by van Fraassen's account. I will use this idea as a starting point and link it with the notion of explanatory hierarchies of Barelbroth and subsequent accounts of unification. As we shall see, this will lead to a resolution of the ad explanandum problem raised by Halonen and Hintikka. It will also help us reach a notion of unification suitable for the evaluation of explanatory nexuses.

Bas van Fraassen's (1980) pragmatic approach to explanations starts with the idea that explanatory questions do not arise in a vacuum, devoid of background information. Instead, they are reducible to contrastive questions of the type Q: "Why x and not y ?", where x is the factual explanandum and y is some contrast class of other unrealized possibilities. Under this condition, an explanation takes the form of an answer A: "It is the case

that x and not y because A ” (for further details and formalization see van Fraassen 1980 and Kitcher 1989). In van Fraassen’s original formulation the answer must (ideally) be a direct answer. A direct answer is one that reveals explanatorily relevant information, which, by its introduction, shows the topic of the question x to be the only factually true statement out of the contrast class y .

Two elements are essential in such an approach to explanations. Firstly, relevance of the explanatory answer to the question and, secondly, the construction of the appropriate contrast class.

A typical example will be a causal explanation in which we have the question “Why did the window shatter when it got hit by the ball, as opposed to the ball just bouncing off it?”, here the answer can contain, say, relevant information about the force of the throw, the weight of the ball, and the tensile strength of the glass. As such, the answer provides case specific conditions (or ad explanandum information) that show why the shattering of the window was the only realized possibility (for similar contextual approach to causation see Schaffer 2005; Reiss 2013a, 2013b).

If we look at this example closely, we will see this to be a case of a very specific question, which receives an equally specific answer; as such it can be treated as an instance of a very stringent pattern. However, under unification we are not typically asking why a singular fact has occurred given any range of possibilities. Instead, we are asking why the range of possibilities is as it is, but also why this fact has occurred under this specific range of possibilities. We are first trying to determine the correct contrast class and then to determine why x has occurred within that class.

In other words, we are not dealing with a singular explanatory seeking question and a singular answer. Instead, we are dealing with a generalization which defines the borders of the contrast class, and then with a series of explanatory seeking questions, nested into each other, that aim to exhaust this contrast class or, more specifically, to zero in on the fact under scrutiny. This nesting works by introducing information that leads to providing a more and more narrow contrast class, until we zero in on the specific singular explanandum. If we take this idea into account, we can easily arrive at the notion of a hierarchy, and also treat this hierarchy as effectively describing an explanatory nexus.

Instead of following Kitcher's approach and defining unified theories statically, we can include more and more factually specific information in our answers. This will permit us to resolve the ad explanandum problem, raised by Halonen and Hintikka. In fact, it will also help us to arrive at the notion of an explanatory nexus and its organization. Similarly, instead of describing nexuses statically as all the available explanations for some facts, we can define clusters of explanations as a series of nested explanatory questions and answers, along with the contrast classes of phenomena which they introduce.

Based on this we can determine the completeness of an explanatory nexus not as all the available explanations for x , but as a sort of recursive interrogative game that includes variations of x and some distinct (but still theoretically relevant) facts y :

Q1: Why $x...x_n$ along with some $y...y_n$ can occur?

A1: $x...x_n$ and $y...y_n$ can occur because A1.

Q2: given A1 why $x...x_n$ occurs and not y ?

A2: $x...x_n$ occurs and not y because A1.

Q3: given A1, A2 why x_k occurs and not x_n ?

A3: x_k occurs and not x_n because A3.

...

For example, we would like to understand why Julius Caesar died. We can start with the question: Was Caesar killed or did he die of natural causes? Given the answer that Caesar was killed, we can ask further: Was he poisoned, stabbed, cut, burned, drowned, suffocated or beaten with a blunt object? Once we establish that Caesar was stabbed we can try to answer further forensic questions about the stabbing: Were any vital organs punctured? Did he die of shock, or because of blood loss, or because of organ failure? Were the lungs or heart punctured?

Such a recursive game will establish a series of nested questions and answers relating more general questions to more particular ones. Two important points emerge. Firstly, in order to make the resulting structure manageable, we must not ask why-questions in any non-restricted way—we should not mix facts about the forensics of the Caesar's death with these of possible motives of his assassins, the political situation in Rome at the time, the psychological dispositions of Brutus etc.

Secondly, even in this restricted fashion, this structure of nested explanations meets Halonen and Hintikka's problem, because at each step we are introducing new information via more specific and fine-grained answers, which in turn permits us to appreciate more and more minute differences between our contrast class of phenomena. Therefore, better understanding intuitively results from grasping not only the basic contrastive difference between some theoretically relevant phenomena x and y , but also more and more minute variations of x .

An interesting problem emerges here. Van Fraassen's original account treated contrast classes unrestrictedly. As a consequence, we can use any contrast class of non-realized possibilities as a basis of our explanation. For instance, our problem of Cesare's death becomes the muddled: Did Cesar ever die or was he abducted by extra-terrestrial time travelers from Mars, who also hypnotized Brutus and other members of the senate, leaving only a dummy dead clone of Cesar on his original place? Therefore, an important problem is how to restrict the contrast class, so we can arrive at a more manageable and realistic structure

As a solution, firstly, we should consider a contrast class that is limited to the actual domain of our explaining theory. Secondly, we should include only genuine scientific explanations, such that the relation between the topic of the question (explanandum) and the answer (explanans) satisfies the requirement of some of the accepted models of explanations. For instance, if we have a structure containing only causal explanations and counterfactual answers, we are not normally including any possible causal influences or all and any variations of the explanandum (See Woodward 2003 and his notion of serious possibilities). Moreover, we can also deploy Kostić's criterion of intimacy between explanandum and explanans as a guideline on how to organize the explanatory structure—e.g., starting with explanations (explanans) with a more immediate relation to the explanandum, and moving towards more complex cases.

Taking these considerations into account we can also quantitatively determine—*How unified the nexus of x can be by solving the problem:*

Prob 1: What is the least number of why questions that need to be answered in order to uniquely pick up x from as rich as possible an initial contrast class?

This problem introduces the requirement for stringency (as informative as possible) and systematization (as few as possible as systematic as possible explanations). Similarly, to the requirement for unification, the smaller the initial contrast class, the easier it will be to arrive at a maximally stringent and exhaustive structure, and vice versa. However, we saw that “grasping” requires evaluating explanations not only in terms of their stringency but also in terms of their broader applicability (systematization).; A such, the larger the contrast class considered, the more explanatory information we would need in order to zero in on the specific explanandum.

Similarly, the *optimal organization and completeness of an explanatory nexus can be determined by the problem:*

Prob 2: What is the least number of why questions that need to be answered in order to pick up each unique member of as rich as possible an initial contrast class?

The central idea is that we are trying first to establish a starting generalization which is as systematic as possible for as large as possible a contrast class of phenomena, and then trying to exhaust this contrast class by the smallest possible number of additional questions that we need to answer in order to exhaust the relative differences between all the phenomena in the contrast class. Such nexuses can be as open or as limited as one requires. For instance, for a complete theory of evolution, we would need an initial generalization like SNS, and then extensions of specific explanations that can ultimately map onto the origin events and population states of all the species on Earth. Obviously, most actual theories and their explanations will have a much narrower scope and can be, for all practical purposes, complete (Newtonian optics or the study of some simple model organisms in biology are good examples).

With this in mind, we can also determine *the subjective completeness of grasp* and the different degrees of understanding that epistemic agent S will have for x by the problem:

Prob 3: How large is the contrast class of which x is a member that S considers, and what is the number of questions that S needs to ask and receive an answer to in order to pick up x uniquely from the contrast class?

Naturally, the less information the agent requires in order to discover why x has occurred as oppose to a large number of variations of x , the better he understands x . Armed with these ideas, we can also represent the resulting structure in a more robust way—as a hierarchical graph tree. By examining its structural properties, I hope to arrive at more precise responses to these problems, and, as such, also at quantitative criteria for evaluating explanatory nexuses.

4.1. Explanatory structures as directed graph trees

A starting point from which to describe these explanatory structures is to represent them as directed graphs. The simple case will be a direct or simple explanation, where A is the explanans and B the explanandum:

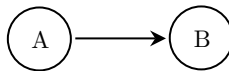


Fig. 1

In order to reach the notion of a pattern, and utilizing van Fraassen's pragmatics, we need only change the structure of explanatory seeking questions from "Why x_1 and not y ?" to the exhaustive Q: "Why x_1 *xor* x_2 ?". Thus, Q along with two possible answers—A1 for x_1 and A2 for x_2 —can be represented as such:

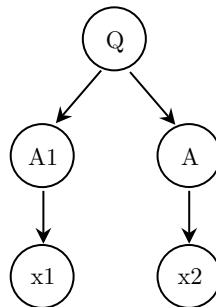


Fig. 2

Because an explanatory pattern is any structure in which the explanans applies to more than one explanandum, we can represent the case from fig

2 as a pattern in which A encompasses both x1 and x2 and instantiates respectively two simple explanations B and C. Recall that the difference between an explanation and a pattern is that a pattern instantiates more than one explanation. Therefore, this will also be the case for the most stringent pattern, which is just a step removed from a singular explanation. In order to represent exclusively the branching and the inter-explanatory relations of such a pattern, we can depict it as a simple binary tree:

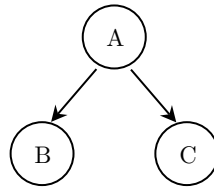


Fig 3.

The paths $A \rightarrow B$ and $A \rightarrow C$ should be read as simple explanations of facts, respectively, x1 and x2; and the structure as covering $A\{x1, x2\}$. Under this approach a root is always the most systematic node. The children are always more stringent and the leaves are always singular explanations that zero in on a particular explanandum. A pattern can cover any range $Rang\{x \in N: 1 < x\}$. Consequently, patterns can be ordered into levels according to their range and nested into each other, where, respectively, patterns of the type $A \subseteq B$ are cases in which A is a more stringent pattern for some of the x that the more systematic B covers. From this perspective, an optimal organization for the structure of such nested explanations will be a balanced binary tree (fig. 4):

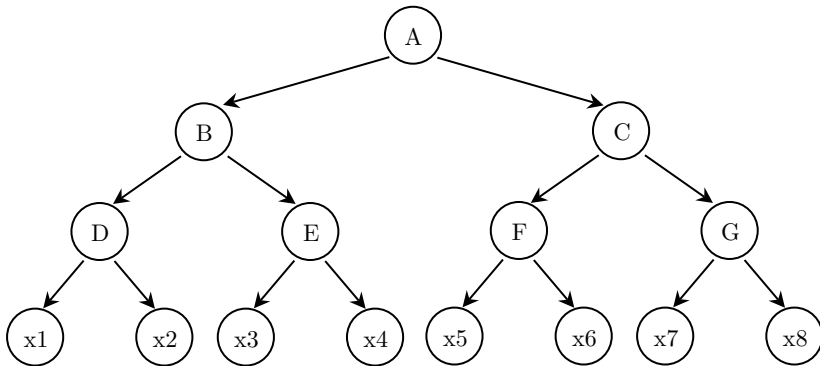


Fig. 4

To be more precise, in such a tree there will be no leaves that directly connect with the root (with the exception of a structure with a single maximally stringent pattern that is not equivalent to a direct explanation—as in fig.3). All the leaves will be connected to a branch and all the branches including the root will have at most 2 children. Because at each level a branch will have at most 2 children, this will be a case of a structure that contains as stringent patterns as possible, that are completely systematic for the leaves. (Recall that the most stringent pattern has only 2 instantiations—see fig. 3).

Another property that we would desire is to keep the level (l) of the structure as low as possible. The level of a node is defined by $1 +$ the number of connections between the node and the root. For example, the root will have level 1, its two children 2, their four children 3 etc. (see fig. 4 for an example). Keeping the path from the root to each specific leaf as short as possible exemplifies the scenario in which we need to solve as few explanatory problems as possible in order to reach each domain specific fact. In other words, we are trying to obtain as stringent as possible and as systematic as possible structure. Therefore, a solution for *Prob 2* (of optimal organization and completeness of an explanatory nexus) essentially reduces to a balancing problem for a binary tree.

The balance of a binary tree can be determined based on the relation between the level of the tree l and the number of its leafs N , which is $2^{(l-2)} < N \leq 2^{(l-1)}$. We can obtain a balanced structure when we have as few levels (l) as possible for N number of leafs under the condition: $\log_2 N + 1 \leq l < \log_2 N + 2$; $l = \lceil (\log_2 N) + 1 \rceil$. It is also important to determine the number of nodes, because they represent our explanatory data. For a binary tree this is easy, since each node can have at most 2 children. Therefore, the number of nodes A can be determined based on the number of leafs N as: $A = 2^{\lceil \log_2 N \rceil} - 1 + N$.

With this in mind, the solution to *Prob 1 of. how unified the nexus is* for a balanced binary tree with root R and $X(x_{i...n})$ number of leafs is simply $\forall x_{i...n} \in R = dx_{i...n}$, where d is the depth of a leaf x (the number of edges we need to travel from the root to the leaf x).

Finally, a solution for *Prob 3—determining the subjective completeness of grasp* will be found in the case of determining what subtree of the

balanced tree A an epistemic subject can recover, or how similar his explanatory nexus is to A . If his explanatory nexus is identical to A , then he completely grasps explanatory structure A . Determining the similarity between such graphs, cannot be only structural but must also contain evaluation of the explanatory data that each node contains. Therefore, determining this similarity is a more complex issue that goes beyond the scope of the present text. However, some suggestions for comparative analyses of data graphs can be found in Zageret al (2008).

Obviously, any actual explanatory structures that we are going to deal with in the wild will fall short of this ideal scenario. They will display varying degrees of balance; they might be incomplete; they may have multiple roots; they may display varying heights in their explananda, and so on. Nevertheless, actual organizations of explanations can be examined as such directed graph trees and their degree of balance be measured. In the next section, I will suggest that, perhaps contrary to initial appearances, such structural analysis is not too idealized, and at least some fragments of theories that use simple nested mathematical models can be represented in this way.

Before moving on to the case study, a final note that I must make is that the approach to unification as a descriptive tool of explanatory nexuses or theoretic explanatory stores in general was inspired by the algorithm problems related with optimal organization of databases for solving problems in obtaining fast and effective searches. As such, the advances made by both computer science (Knuth, 1998) and formal analyses of such structures from descriptive set theory (Kechris, 1994), might offer interesting avenues of research.

4.2. Case study – ecological models as explanatory structures

To start building on the idea of a theory as a hierarchy of explanations, I will use an example from the ecological theory of predation. I believe, however, that similar analysis can be undertaken for any theory which has a clearly definable conceptual core and displays sufficient typicality in its explanatory strategies.

Even though the complexity of ecological phenomena does not permit the same level of robust unification as displayed by law-based theories in physics, reconstructing the ecological theory of predation as a unified system is not a philosophical chimera. Some ecologists (Holt 2011; Ginzburg and Colyvan 2004, p.135) have already attempted such reconstructions and the present case study is based on their work.

My starting assumption is that the ecological theory of predation can be partially ordered as a hierarchy of models, where more specific models are nested into more general ones. The resulting hierarchical structure exhausts a significant fragment of the factual domain and its models are indispensable for explaining population dynamics. This structure of related models can also be examined as an organization of explanatory patterns. The resulting structural representation can provide valuable information on how ecological explanations offer understanding. But before proceeding onward, some preliminary clarifications must be made.

Firstly, proponents of the idea that scientific theories are collections of loosely related models might object that models are abstract or concrete objects and as such cannot be completely reduced to propositions, whilst the analysis developed here is exclusively concerned with explanations which are patently propositional. Without getting into the thorny problem of scientific representations, we can assume that models express beliefs about their targets, and that models are employed to make inferences about their targets. Consequently, models can be loosely investigated as sets of functions which interpret as parameters some of the core theoretic terms and predicates of the theory. Given an input that substitutes the parameters with values derived from observations, these functions will generate conclusion-like statements about observational targets. If the model manages to be representative of its targets, the conclusions made by the model, along with its fundamental assumptions, can be ordered in an explanatory inferential structure.⁹ Secondly, mixing models and explanatory patterns can result in a somewhat messy picture of theories. Nevertheless, I will try to

⁹ This claim does not aim to diminish the significance of the philosophical theories of representation, nor to completely reduce the semantic view of theories to a syntactic one. My modest goal here is to establish that there is an intimate relation between explanatory patterns and formal models.

present a clear picture as best as I can by giving examples of how explanatory patterns can be used to derive formal models and vice versa.

Any ecological theory starts with the assumption that the demographics of a population will depend on 3 factors: i.) *E* environmental conditions (for instance available resources, fluctuations of the climate etc); ii.) *N* factors specific for the organisms (as birth and death rates); iii.) *P* factors due to interaction with other organisms (such as predation, competition for resources, parasitism, cooperation etc.). This very general framework of ecological relations can be specified as a root pattern:

GenEc:

Explanans: The influence of species-specific factors N', environmental factors E and relations with other species P → Explanandum: determines the demographics of a population of organisms N at a time t.

Based on it we can also build an initial mathematical expression (based on Holt 2011):

$$(eq. 1) \quad \frac{dN}{dt} = Nf_n(N,P,E)$$

As in our toy case concerning natural selection, such a pattern, and the resulting model, are not very informative. *GenEc*, however, serves as a root pattern that draws the boundaries of the theory. It has only the systematizing function of drawing the boundaries of the domain of ecology. Such a pattern does not provide any contrast between the domain-specific facts. Consequently, it can only provide understanding if it is extended to more informative explanatory patterns that specify its core concepts (*E*, *P* and *N*). A step in that direction is to specify only one of the conditions and examine the resulting contrast that such a pattern can provide.

The most straightforward way to extend one of the factors is to assume a function of birth-rate for *N* and construct a model of population growth:

$$(eq. 2) \quad \frac{1}{N} \frac{dN}{dt} = r_N$$

The solution of this simple function is exponential growth: $N(t) = N_0 \exp(r_n t)$. To represent it as a pattern we can interpret the function as receiving input information about a population of organisms at time t^0 and

generating output populations states for t^n . The model then gives rise to the pattern:

ExpG:

Explanans: The unbound exponential birth rate of a population of organisms N at time $t^0 \rightarrow$ Explanandum: determines the demographics of the population N at time later time t^n .

It might seem that such a pattern is just as uninformative as our starting generalization of ecology (GenEc). However, we should recall that we are not required to evaluate this pattern and its model in isolation, but as a branch of the initial generalization. As such this model should be more contrastive than GenEc. The exponential growth pattern offers understanding only as an extension of the general assumptions of GenEc; by specifying only one of the factors (that of birth rate) it effectively negates environmental and inter-species factors. This creates a contrast with all the situations in which such factors have a noticeable effect on the population. Interestingly, even this simple expression of exponential growth has some domain of application. In microbiology exponential growth is used to describe microorganisms' growth until nutrient exhaustion (Slavov et al 2014). In epidemiology, it is employed as a starting point for explaining pandemic diseases (Chowell and Viboud 2016; Brauer and Castillo-Chavez 2012). Finally, in ecology, it is used for describing the behaviour of the prey when there are sufficient nutrients and an absence of predators (Arditi & Ginzburg 2012).

Returning to our initial structure, we can extend the root pattern (GenEc), by introducing a relation with another species. This brings about the model:

$$(eq. 3) \quad \frac{dP}{dt} = Pf_p(P, N, E)$$

This new equation stands on the same level of generality as our initial equation (1), because it simply introduces another population of organisms. Similarly, it can only be made more precise if the two populations are coupled. A straightforward coupling of equation (1) and (3) would be a case of predation, where P is a predator and N prey. This results in the pattern of basic predation:

BasicPred:

Explanans: If a population P's growth depends on a resource provided by the member of a population N → Explanandum: then the harvesting of N will hamper the growth of N and benefit the growth of P.

In this case we can assume that the resources that P depends on are the members of N, and that harvesting directly diminishes the population of N and benefits P. Adding our previous notion of exponential growth of N and an extra assumption of the death rate of P, leads us to the Lotka-Volterra equations (Volterra 1926):

$$\frac{dN}{dt} = rN - aNP \quad (4.1)$$

$$\frac{dP}{dt} = eaNP - qP \quad (4.2)$$

Here N and P are the demographics of prey and predators, respectively; rN and qP designate the birth rate of the prey and the death rate of the predators, respectively; e represents the efficiency of converting prey biomass to predator biomass; and a the searching efficiency or the attack rate of predators. Of central importance for these equations are the so-called numerical and functional responses. The functional response is the intake rate of the predators as a function of the density of the prey. Here it is represented by the kill rate aNP . The numerical response is the change in predator density as a function of change in prey density; it is given in equation 3.2 by a function of the kill rate aNP that depends only on the efficiency e of the conversion of prey biomass to predator biomass.

This short overview of the Lotka-Volterra equations shows that they introduce quite a few new details into our hierarchical organization of ecological modelling. Although the equations remain highly idealized, they nevertheless serve as a generalization by which we can distinguish between the three possible population states of a coupled predator-prey population—ecological balance, double extinction and the extinction of the predator. As such the specification of the variables of these models and addition of specific parameters can be used to describe (or zero in to) a specific population state or dynamics (Abrams & Ginzburg 2000; Petkov 2019). If this state is representative of some targeted ecological populations, the model can also

be used to derive specific explanations as to why such a state has been observed.

Without going into further details by describing more specific models, ones that have different functional responses, or take into account the carrying capacity of the environment or the satiation of the predators, we can organize the equations introduced so far into a structural hierarchy and investigate its *properties*. These ecological models can be given as the following preliminary tree:

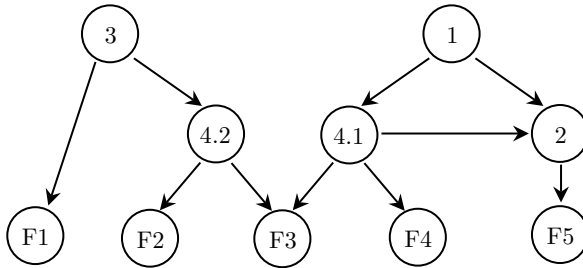


Fig.5

In this graph, the numbered nodes correspond to the equations (1 to 4.2) and the leaves (F1 to F5) can be assumed to represent specific explanations that can be derived from these models. Here, these specific explanations zero in not on singular facts but instead on the type of facts that they cover. As mentioned earlier, by adding more specific parameters we can use these models to describe and explain specific ecological facts. Although the resulting picture is significantly simplified, it nevertheless illustrates that F5 can stand for populations that can be explained by the exponential growth model (2.); F3 can represent facts related to an equilibrium between predator and prey; F4, facts related to population dynamics, in the absence of predators but considering limiting environmental factors; F2, a situation of prey extinction; and F1, a situation in which the other species is also coupled with resources other than the prey. It also shows that F1 is a type of population states that yet lack specific descriptions within our theory. Importantly the structure also designates that there is a relation between equations (4.1.) and (2.), as, in the absence of predators, the equation (4.1) collapses to the earlier equation of exponential growth (2.). (For

further analysis on the relation between exponential growth models and the Lotka-Volterra model see Arditì and Ginzburg 2012).

We can appreciate that the emerging ecological theory of predator-prey dynamics is describable as a balanced tree and that it can be treated as an example of a unified explanatory nexus that also presents a complete explanatory structure, at least for the basic states of predator-prey populations—unlimited growth, ecological balance, double extinction and the extinction of the predator. As such, this hierarchy of models can serve as a good illustration of a theory that satisfies the ideas of optimal organization of inferential explanatory patterns in terms of their relations. Consequently, such a structure can be used to also describe the understanding of specific population states, as contrasted with other possible scenarios for population dynamics. For instance, the discipline of conservation biology requires such contrast, since it aims at maintaining and explaining biodiversity by defining the conditions under which an ecological system maintains balance, as opposed to various extinction scenarios.

5. Conclusion

Understanding of phenomena via an explaining theory requires both the right type of epistemic agents and the right kind of theory. As such understanding has two complementary dimensions—a subjective and an objective one. Whilst under the concept of “grasping” the analysis of the subjective dimension has undoubtedly shown that explanations cannot be considered in isolation, the analysis of the objective dimension has been mostly preoccupied only with uncovering the internal inferential relations within singular explanations. Criteria for evaluating the specific structure and organization of explanatory clusters or nexuses has received much less attention. The present study aimed at filling this gap.

As we saw the particular features of the explanatory inference and the position of the explanations within a theoretic hierarchy of explanations are equally important objective criteria for evaluating the resulting explanatory understanding. Moreover, the position of particular explanations within such structures can potentially amend some of these explanations’ shortcomings as for instance if more general and less informative explanations

instantiate more stringent ones. In this short essay, I represented these structures as directed graph trees and thus aimed at providing more robust criteria for evaluating the degree of understanding that such explanatory structures can offer in terms of completeness and balance.

The short exposition on ecological modeling serves as further evidence that such structural hierarchies are not a purely abstract idea and at least some fragments of scientific theories can be represented as such.

Funding

This work is supported by China's Major National Social Science Project—19ZDA041.

References

- Abrams, Peter A. and Ginzburg, Lev R. 2000. "The Nature of Predation: Prey-Dependent, Ratio-Dependent or Neither?" *Trends in Ecology and Evolution* 15 (8): 337–41. [https://doi.org/10.1016/S0169-5347\(00\)01908-X](https://doi.org/10.1016/S0169-5347(00)01908-X)
- Arditi, Roger and Ginzburg, Lev R. 2012. *How Species Interact: Altering the Standard View on Trophic Ecology*. New York: Oxford University Press.
- Bartelborth, Thomas. 2002. "Explanatory Unification." *Synthese* 130 (1): 91–107. <https://doi.org/10.1023/A:1013827209894>
- Baumberger, Christoph, Beisbart, Claus, and Brun Georg. 2017. "What is Understanding? An Overview of the Recent Debates in Epistemology and Philosophy of Science." In *Explaining Understanding: New Perspectives from Epistemology and the Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon, 1–34. New York: Routledge.
- De Regt, Henk W. 2015. "Scientific Understanding: Truth or Dare?" *Synthese* 192 (12): 3781–797. <https://doi.org/10.1007/s11229-014-0538-7>
- Dowe, Phil. 2000. *Physical Causation*. New York: Cambridge University Press.
- Friedman, Michael. 1974. "Explanation and Scientific Understanding." *Journal of Philosophy* 71 (1): 5–19. <https://doi.org/10.2307/2024924>
- Ginzburg, Lev R. and Colyvan, Mark. 2004. *Ecological Orbits: How Planets Move and Populations Grow*. New York: Oxford University Press.
- Grimm, Stephen R. 2010. "The Goal of Explanation." *Studies in the History and Philosophy of Science* 41 (4): 337–44. <https://doi.org/10.1016/j.shpsa.2010.10.006>
- Gurova, Lilia. 2017. "On Some Non-Trivial Implication of the View that Good Explanations Increase Our Understanding of Explained Phenomena." *Balkan Journal of Philosophy* 9 (1): 45–52. <https://doi.org/10.5840/bjp2017914>

- Halonen, Ilpo and Hintikka, Jaakko. 1999. "Unification—It's Magnificent but is it Explanation?" *Synthese* 120 (1): 27–47.
<https://doi.org/10.1023/A:1005202403274>
- Hempel, Carl G. 1965. "Aspects of Scientific Explanation." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 331–96. New York: Free Press.
- Holt, Robert D. 2011. "Natural Enemy-Victim Interactions: Do We Have a Unified Theory Yet?" In *The Theory of Ecology*, edited by Samuel M. Scheiner and Michael R. Willig, 125–61. Chicago: University of Chicago Press.
- Kechris, Alexander S. 1994. *Classical Descriptive Set Theory*. Dordrecht: Springer.
<https://doi.org/10.1007/978-1-4612-4190-4>
- Khalifa, Kareem. 2017. *Understanding, Explanation, and Scientific Knowledge*. Cambridge, UK: Cambridge University Press.
- Kitcher, Philip. 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, edited by Philip Kitcher and Wesley Salmon, 410–505. Minneapolis: University of Minnesota Press.
- Knuth, Donald. 1998. *The Art of Computer Programming. Volume 3: Sorting and Searching*. Second Edition. Massachusetts: Addison-Wesley.
- Kostić, Daniel. 2018. "Minimal Structure Explanations, Scientific Understanding and Explanatory Depth." *Perspectives on Science* 27 (1): 48–67.
https://doi.org/10.1162/posc_a_00299
- Lange, Marc. 2016. *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. New York: Oxford University Press.
- Lewis, David. 2000. "Causation as Influence." *Journal of Philosophy* 97 (4): 182–97. <https://doi.org/10.2307/2678389>
- Morrison, Margaret. 2000. *Unifying Scientific Theories*. Cambridge: Cambridge University Press.
- Newman, Mark. 2014. "EMU and Inference: What the Explanatory Model of Scientific Understanding Ignores." *European Journal for Philosophy of Science* 4 (1): 55–74.
- Salmon, Wesley. 1998. *Causality and Explanation*. New York: Oxford University Press.
- Schaffer, Jonathan. 2005. "Contrastive Causation." *Philosophical Review* 114 (3): 327–58. <https://doi.org/10.1215/00318108-114-3-327>
- Petkov, Stefan. 2015. "Explanatory Unification and Conceptualization." *Synthese* 192 (12): 3695–717. <https://doi.org/10.1007/s11229-015-0716-2>
- Petkov, Stefan. 2019. "Studying Controversies: Unification, Contradiction, Integration." *Journal for General Philosophy of Science / Zeitschrift für Allgemeine Wissenschaftstheorie* 50 (1): 103–28. <https://doi.org/10.1007/s10838-018-9431-2>

-
- Reiss, Julian. 2013a. "Contextualising Causation Part I." *Philosophy Compass* 8 (11): 1066–75. <https://doi.org/10.1111/phc3.12074>
- Reiss, Julian. 2013b. "Contextualising Causation Part II." *Philosophy Compass* 8 (11): 1076–90. <https://doi.org/10.1111/phc3.12073>
- Van Fraassen, B. 1980. *The Scientific Image*. Oxford: Clarendon Press.
- Volterra, Vito. 1926. "Fluctuations in the Abundance of a Species Considered Mathematically." *Nature* 118: 558–60. <https://doi.org/10.1038/118558a0>
- Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press. <https://doi.org/10.1111/j.1933-1592.2007.00012.x>
- Zager, Laura A. and Verghese, George C. 2008. "Graph Similarity Scoring and Matching." *Applied Mathematics Letters* 21 (1): 86–94. <https://doi.org/10.1016/j.aml.2007.01.006>

How Not to Argue about the Compatibility of Predictive Processing and 4E Cognition

Yavuz Recep Başıoğlu*

Received: 16 February 2020 / Revised: 26 May 2020 / Accepted: 9 June 2020

Abstract: In theories of cognition, 4E approaches to cognition are seen to refrain from employing robust representations in contrast to Predictive Process, where such posits are utilized extensively. Despite this notable dissimilarity with regard to posits they employ in explaining certain cognitive phenomena, it has been repeatedly argued that they are in fact compatible. As one may expect, these arguments mostly end up contending either that Predictive Process is actually nonrepresentational or that 4E approaches are representational. In this paper, I will argue that such arguments are inadequate for the indicated purpose for several reasons: the variety of representational posits in Predictive Process, the diverse attitudes of practitioners of 4E approaches toward representations and the unconstrained use of the term “representation” in cognitive science. Hence, here I will try to demonstrate that any single argument, if it depends on representational 4E approaches or nonrepresentational Predictive Process, falls short of encompassing this heterogeneity in pertinent debates. Then, I will analyze similar arguments provided by Jacob Hohwy and Michael Kirchhoff to illustrate how destructive this seemingly ordinary criticism is.

* University of Osnabrück

 <https://orcid.org/0000-0003-4966-1144>

 Institute of Cognitive Science, University of Osnabrück, Wachsbleiche 27 49069, Osnabrück, Germany

 basogluyavuz@gmail.com

© The Author. Journal compilation © The Editorial Board, *Organon F*.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0).

Keywords: 4E cognition, embodied cognition, free-energy principle, mental representation, predictive processing, representation wars.

1. Introduction

While 4E approaches to cognition (i.e., embodied, extended, embedded, and enactive; one might also include ‘situated,’ ‘distributed,’ ‘affective,’ ‘interactive,’ ‘extensive,’ etc.) (4E cognition henceforth), which are also gathered under the generic name “second-generation cognitive science” (Lakoff and Johnson 1999, 78), have enjoyed their heyday over the last decades, another idea has also recently started to excite the cognitive scientist: Predictive Processing (PP henceforth) as a particularly promising neuro-computational framework purporting to account for a uniform understanding of the mechanisms underlying cognition.

Whereas practitioners of 4E cognition are rigorously at odds with traditional cognitivists who recruit internal symbolic representations ubiquitously in their explanations of cognitive phenomena, PP is thought to make use of alleged representations nontrivially to use an internal model of the external world. For this reason, *prima facie*, these two theories of cognition seem to be incompatible. That is, at least one must be false. Nevertheless, philosophers of mind and cognitive scientists have been at pains to prove that PP and 4E cognition are compatible (e.g., Clark 2013; 2015a; Kirchhoff 2018; Gładziejewski 2017; Hohwy 2018). Hence, to establish the compatibility of PP and 4E cognition in terms of representations, scholars mostly argue either that 4E is representational or that PP is nonrepresentational to leave no fundamental difference that might suggest that these theories of cognition are, in fact, incompatible.

In section 2 of this paper, I first provide a brief and selective overview of PP and 4E cognition and expound how the question of their compatibility boils down to questions of the ontological nature of representational posits deployed by these theories. In section 3, I try to explicate the general schema of arguments advanced when arguing for the compatibility of PP and 4E cognition in terms of representations. I shall argue that any argument contending that PP and 4E cognition are compatible is doomed to fail when it depends on either representational 4E cognition or nonrepresentational PP.

This is the case because the term “representation” as employed by the cognitive scientist, the attitudes of 4E cognition proponents toward representations, and the kinds of representational posits applied in the PP framework are not homogenous. It is thus not possible for a single argument concerning the existence of representations used in any of these theories to embrace this heterogeneity. It is no longer obscure that the unconstrained use of the term representation creates a flexibility to read any theory in representational or nonrepresentational terms. This trivializes the representationalism vs. anti-representationalism debate and renders them “for the sake of appearance” (Haselager et al. 2003, 21). This, I shall argue, turns out to be more daunting in debates on the relation between 4E cognition and PP. In section 4, I analyze the arguments of both Hohwy (2018) and Kirchhoff (2018) for the compatibility of PP and 4E cognition and show that both arguments employ the flawed argument schema scrutinized in section 3. Thereafter, the main task of this section is to illustrate how Hohwy and Kirchhoff’s arguments fail to be conclusive in establishing the compatibility of PP and 4E cognition. This illustration aims to emphasize how significant the seemingly simple criticism voiced in section 3 is and how destructive it can be when ignored.

2. 4E cognition, predictive processing and the representation wars

Ushering in a whole new era in the philosophy of mind and cognitive science, 4E cognition arose as a result of growing discontent with traditional cognitivism’s claim that cognition is brain-bound or intracranial and is a kind of computation realized by syntactically manipulating symbolic mental representations (see Fodor 1975 and Pylyshyn 1984). Lakoff and Johnson’s (1980) idea of the metaphoric use of language as not peripheral to cognition but as fundamental to how one conceptualizes the world, including time, space, feelings, etc., and Varela et al.’s (1991) emphasis on embodied action especially initiated this approach. New studies and developments in robotics (e.g., Brooks 1991) and dynamical systems (e.g., Thelen and Smith 1994) have also accelerated the advent of 4E cognition. While it seems quite futile

to attempt to provide a uniform definition encompassing the whole research program for 4E cognition as is made evident in section 3, its most distinctive characteristics can be given as follows: it prioritizes the idea of the body¹ and environment as *constitutive* in cognition rather than as *just peripheral* to certain cognitive processes. This means that in favor of an embodied and extended approach, 4E cognition undermines the intracranialism (e.g., Adams and Aizawa 2008) associated with traditional cognitivism, which asserts that cognition, as a matter of empirical fact, is brain-bound or intracranial. 4E cognition also accentuates nonlinear constant dynamical coupling between action and perception, and among the brain, body and environment, which purports that intelligent behavior is not driven by a certain form of computation but results from this dynamical coupling. Thus, the body and the environment not only causally contribute to cognition but also play a constitutive role, and the brain, the body, and the environment are not separated from each other, but constitute one continuously interacting system.

Since the body and environment have constitutive roles in cognition and are in constant dynamical coupling, they are in a position to provide the information required for cognition *on the fly*. 4E cognition thus defends the notion that representations that are supposed, according to the cognitivist, to carry such information are *mostly* not needed nor desirable in explaining cognitive phenomena. Thereupon, most importantly for the issue at hand, 4E cognition researchers (e.g., Goldman 2012; Van Gelder 1995; Noë 2004) aim to minimize if not altogether to dispense with the elementary role of amodal, action-neutral, quasi-linguistic, contentful, and symbolic inner representations of cognitivist understandings of the mind. This tendency among practitioners of 4E cognition quickly escalated ‘the representation wars’ between representationalists and anti-representationalists, and even after almost three decades of discussion, these representation wars have not yet been settled.

In the thick of the war, PP posits that to engage the world around it, the brain,² which uses Bayesian reasoning, continually predicts incoming

¹ The term “body” refers to extracranial parts of the body or to the body minus the brain in this context.

² The words ‘mind,’ ‘brain,’ and ‘agent’ in this discussion appear to be synonymous. Their use is a matter of taste. Slightly different connotations relative to their

sensory data by means of its top-down hierarchical architecture, which models hidden nested causal complexities of the external world. The predictions generated by these top-down processes are compared to incoming sensory data, and error messages are generated when there is a mismatch between top-down predictions and the incoming sensory data. Generated error messages propagate up the hierarchy to adjust the internal model of the environment, which in turn helps minimize errors in predictions made at lower levels of the hierarchy and hence generates more accurate predictions for following iterations. To minimize prediction errors, the internal model of nested causal complexities of the environment can be updated and adjusted as just explained. The brain's other strategy in performing prediction error minimization (PEM henceforth) involves moving its sensory organs in a certain way to fit incoming sensory data to the prediction already generated by the system's hierarchical architecture. In so doing, the brain proactively samples sensory data. This particular mode of PEM is called *active inference*, which emphasizes continuous dynamical coupling between perception and action and between the environment and the agent. Thus, "perceiving and acting are but two different ways of doing the same thing" (Hohwy 2013, 76).

To appreciate the peculiarities of PEM, one must also place it in the context of the Free Energy Principle (FEP henceforth), according to which "any self-organizing system that is at equilibrium with its environment must minimize its free energy" (Friston 2010, 127). This means that an organism strives to maintain homeostasis in the face of continuously changing environmental factors. The organism avoids situations that are unexpected or surprising (or more technically showing higher levels of *surprisal*) for its phenotype. In rendering the PEM strategy of PP a special version of the FEP, top-down predictions are generated in accordance with bodily and environmental states, which are expected not to exhibit higher surprisal relative to its phenotype (e.g., being out of water has high surprisal relative to the fish phenotype but not to the human phenotype, and accordingly, due to the difference between their phenotypes, the fish expects to be in water, whereas the human does not.) This suggests that "all aspects of perception and

original fields are ignorable for the issue at hand. Throughout the paper, I respect the choices of the authors in question and use them accordingly.

cognition then have a foundation in bodily states and movement and purposeful behavior have a foundation in the environment” (Hohwy 2018, 137).

The PEM strategy is then what underlies cognition. In pursuing PEM, the brain can detect the hidden causes that best explain sensory data impinging upon its sensory organs. Thus, the brain is nothing but a continuously working prediction machine (see also Clark 2013, 2016).

Due to the scope of this paper, several aspects of the theories in question are intentionally ignored. However, for the present study, this selective overview of PP and 4E cognition shall demonstrate that proponents of PP have made certain cases to claim that it is compatible with 4E cognition on the grounds that both emphasize dynamical coupling among the brain, body, and environment and give fundamental roles to the body and environment in cognition. “There remains, however, at least one famously vexed issue” (Clark 2016, 291): the issue of representation. Nevertheless, due to its aforementioned 4E-friendly tenets, Clark (2015a) (see also Madary 2015) was expecting PP to bring about peace and to end the representation wars. However, it instead opened a new front.

Alleged representations of the PP framework have been, and still are, central to various discussions, and especially to those regarding the relation between PP and 4E cognition. Two different lines of discussion are still going on: whether representations are involved in cognition and whether representational posits of PP are robust representations. The jury is still out on both discussions. For the latter, among the camps are conservative and radical versions of PP. While supporters of conservative PP argue that what PP construes are robust representations that are employed in a non-trivial way (see Gładziejewski 2015; Hohwy 2014), radical PP argues for a somehow deflated version of representations of PP. Radical PP thus claims either that representational posits in PP are not full-blooded or robust representations like those embedded in cognitivism (Kirchhoff and Robertson 2018) or that at least not all representational posits are robust representations of cognitivist understanding (Clark 2015b; Orlandi 2014). Generally, discussion on the relation between PP and 4E cognition has considered the above division: conservative and radical PP. Interestingly, however, a great many in both camps argue that their version of PP is 4E cognition friendly (e.g., Gładziejewski 2017; Clark 2016; Kirchhoff 2018; Hohwy 2018).

3. Representational 4E cognition versus nonrepresentational PP

In arguing for or against the compatibility of PP and 4E cognition in terms of representational posits, the typical strategy adopted is as follows. First, one demonstrates the tenets achieved by the PP framework, which match the characteristics of 4E cognition such as those briefly overviewed in section 2. However, this forces her to make one last decision between a representational reading or nonrepresentational reading of both theories, since while one theory uses explicitly representational terminology, the other abstains from such terminology as much as it can. Thus, second, she focuses on certain properties (e.g., decoupleability) of the representational posits, which are called on to perform 4E cognition-friendly tenets and to determine whether such properties qualify them as robust representations depending on a specific definition such as Ramsey's (2007). Through this line of reasoning, one discusses the existence and employment of representations. Thus, the following is the argument schema of this way of arguing:

(Non) Representationality Argument:

- 1) With its means and strategy, PP is able to realize 4E cognition-relevant cognitive tenets.
- 2) PP is not representational (or 4E cognition is representational) because representational posits needed to embrace these tenets of 4E cognition are (or are not) robust representations according to such a definition.
- 3) Therefore, PP and 4E cognition are compatible.

I shall argue that any (non) representationality argument is inefficient because we do not have a uniform idea of representationalism or anti-representationalism to argue that the second premise is true. The need for such uniformity and the problems caused by a lack of it have already been repeatedly pointed out in this debate (Haselager et al. 2003; Svensson and Ziemke 2005; Ramsey 2007). In what follows, it will be shown that this problem is more virulent than once thought for the issue at hand and in fact renders all (non) representationality arguments for the compatibility of PP and 4E cognition invalid.

The first reason relates to the fact that the ontological nature and representational status of posits of the PP framework are still highly controversial even among the most prominent proponents of PP. As to the status of representational posits of the PP framework, most involved in this discussion, whether conservative or radical, accept Ramsey's "job description challenge" as the criterion to be a robust representation.³ This challenge aims to constrain the liberal use of this term and to prevent the term from being trivialized. (Ramsey 2007) asks philosophers if the internal states they wish to call "representations" satisfy this challenge. If they do not, Ramsey suggests that they are not representations proper. Although job description challenge is a theoretical term, it depends on the daily use of the term "representation": to describe the representational function of representational posits in a cognitive system (Ramsey 2007). Thus, theories that posit representations are under an obligation to genuinely demonstrate the sense in which these internal states function as representations or as *stand-ins* for the external state of affairs. In his own words, he asks for "a job description that tells us what it is for something to function as a representation in a physical system" (Ramsey 2007, 27). He argues that posits of new generation theories in the sciences of mind (such as Dretske 1988) fail to meet this challenge and do not qualify as representations, but he does not comment on PP's representational posits.

As the first aspect of the problem, proponents of PP who accept Ramsey's definition as the criterion to qualify as a representation conclude different results. On one hand, some (e.g., Gładziejewski 2015; Kiefer and Hohwy 2017) argue that according to Ramsey's definition, the representational posits of PP satisfy this challenge and serve as structural representations. Gładziejewski (2015), for example, employs a compare-to-prototype strategy also offered by Ramsey as a way to describe the role of internal states. Through this strategy, one finds a structure in daily life referred to as representation in an uncontroversial way and finds that this structure functions in a similar way to the internal states that one wishes to call representation. Gładziejewski argues that internal states in PP resemble cartographic maps in their job descriptions. Thus, he concludes that such

³ I would like to thank the anonymous referee for pressing me on this point.

posits are “(1) structural representations that (2) guide the actions of their users, (3) do so in a detachable way and (4) allow their users to detect representational errors” (Gładziejewski 2015, 567). On the other hand, some believe that representational posits in PP cannot be called representations (e.g., Kirchoff and Robertson 2018, Orlandi 2014). Proponents of radical PP draw attention to generative models of the lower-level domains (e.g., perception) and claim that such architectures are, as it were, model-free and incapable of, say, detaching from what they stand for. One reason for this disagreement relates to the fact that radical and conservative PP focuses on different parts of the system, which will be clearer in the next section. As the second aspect, for this reason, some (e.g., Dolega 2017) argue that Ramsey’s challenge is inadequate for PP that makes use various generative models with a multilayered architecture. With this challenge, Ramsey seems to tie representations to their functions. In his treatment, among the many other properties, the function of internal states emerges as the decisive property. Such complex structures and components of PP allow for different functional readings as seen. Consequently, Dolega rejects Ramsey’s criterion because it is not a suitable challenge for PP’s complex and multilayered architecture, arguing that the status of such posits should be discussed while considering the content rather than representational functions.

Thus, this discussion of job description challenge reveals that, let alone discussions of whether PP is representational, there is no consensus on some prior points: (1) which properties of representational posits should be central to the evaluation of generative models of PP, (2) whether the job description challenge is suitable for PP’s generative model, and (3) which parts of the system should be considered when evaluating PP. While nothing is for sure, the fact is that these representational posits of the PP framework are not the same as those of the cognitivist tradition. There are still broad disagreements about which definition is to be applied to the representational posits of PP and whether they are robust representations according to any given definition.

In addition to ambiguities as to how to evaluate the generative model, there is more than one type of posit of PP that might be considered a robust representation. As Gładziejewski (2017) rightly appreciates, in PP there are at least four more types of posits that might also earn the status of robust

representation: the generative model, sensory signals, prediction error signals, and precision estimators. Orlandi (2014) argues that other posits such as prediction signals, and error signals do not satisfy Ramsey's job description challenge. In debates on representational posits of PP, theorists generally only mention the generative model while ignoring the other posits, which also deserve closer scrutiny and might be decisive in this debate.

The most significant issue anyone discussing the compatibility of 4E cognition and PP in terms of representations must keep in mind is that the term 'representation *as rejected* by the proponents of 4E cognition' is not a homogenous term. It comes with two aspects. As the first aspect, 4E cognition theorists argue against representations on the basis of different definitions and take distinct attitudes toward them. As the second aspect, proponents of 4E cognition raise various challenges against representations.⁴ For instance, Chemero (2009) argues against representations on the basis of Millikan-style representations (Millikan 1984; 1993), which he thinks carry the general characteristics of traditionally employed robust representations. He argues that action-oriented (Clark 1997), pushmi-pullyu (Millikan 1995), indexical-functional (Agre and Chapman 1987) and emulator representations (Grush 1997; 2004) are robust representations. In contrast, Gallagher (2008; 2017) renounces the representational status of these representations on the basis of Rowlands' definition (Rowlands 2006), which interestingly echoes Millikan-style representations. Surprisingly, Thompson, coauthor of *The Embodied Mind* (Varela et al. 1991), which is regarded the *urtext* of 4E cognition, expressed his sympathy for the emulator account of mental imaginary in (Foglia and Grush 2011), which makes use of emulator representations (Grush 1997; 2004). He wrote that "[I] see it as a friendly supplement to my remarks about sensorimotor processes in mental imagery (Thompson 2011, 194)" while granting its status as a form of representation. Thus, these three thoroughgoing proponents of 4E cognition have completely disparate attitudes toward a certain sort of representational posit.

For the second aspect, Varela et al. (1991) argue against the traditional cognitive function of the mind. They suggest that the function of cognition

⁴ I by no means aim to be exhaustive in revealing the diversity of versions of 4E cognition. It suffices for the present purposes to show how many different versions there could be.

is to engage the world and not to think or build an action-neutral duplicate of the world. Accordingly, they are against the traditional function of representations of creating an action-neutral internal model of the external world. Barsalou (1999; 2008) points out what is called “the symbol grounding problem” and rejects the amodality of representations of traditional cognitivist understanding. Goldman (2012) proposes a somehow weakened version of embodied cognition and accepts what he has dubbed “bodily formatted representation,” whose content is information about bodily states collected by means of interoception, and for him, this is how the body earns its constitutive role in cognition. Hutto and Myin (2013; 2018) raise the content determination problem (or what they call “the hard problem of content”), which refers to the impossibility of physical states (in this case, representations) happening to bear any content whatsoever. Thus, what they reject is to employ any content-bearing representational posit.

Hence, any work not speaking to the concerns mentioned above is doomed to be inconclusive, as neither “representation” nor anti-representationalism is homogenous. Thence, to what extent representational posits of PP satisfy the various criticisms raised by proponents of 4E cognition should be central to discussions of the compatibility of PP and 4E cognition. Any (non) representationality argument falls short of revealing this relation. Whenever one accepts a definition to argue for the (non) existence of representation in PP or 4E cognition, there will always be some equally consistent counter definitions that might be employed against him. Whenever one argues for or against the compatibility of PP and 4E cognition on the basis of the (non) existence of representations in either PP or 4E cognition, one will be arguing against or for only a specific version of 4E cognition or anti-representationalism.

It has been already argued that “[w]ithout a properly constrained notion of representation, the debate between representationalists and anti-representationalists is bound to remain a debate for the sake of appearance” (Haselager et al. 2003, 21) because the liberal and unconstrained use of representation as a term in both camps makes any debate between them futile. In this section, I hope to have shown that in the case of representation wars between PP and 4E cognition, the problem is deeper and more fundamental than in any representationalism vs. anti-representationalism debate.

In what follows, I aim to illustrate how destructive this criticism is by analyzing arguments for the compatibility of PP and 4E cognition provided in (Hohwy 2018) and (Kirchhoff 2018).

4. An analysis of Hohwy and Kirchhoff's arguments

The Oxford Handbook of 4E Cognition (Newen et al. eds. 2018) contains two seminal articles arguing that PP and 4E cognition are indeed compatible. In chapter 7 of this book, Hohwy (2018) spells out how PP, about which he has been theorizing (e.g., Hohwy 2013; 2014; 2017), is congenial to 4E cognition “only because 4E cognition, rightly understood, is nothing but representation and inference” (Hohwy 2018, 130). In chapter 12, Kirchhoff (2018) contends that PP is compatible with the nonrepresentational thesis of 4E cognition.⁵ Thus, both argue that PP and 4E cognition are congenial, but for different reasons. One suggests that 4E cognition is representational while the other proposes that PP is not representational. I shall argue here that both fail to be conclusive. In (Hohwy 2018) and (Kirchhoff 2018), the term “representation” is used ambiguously, and they do not explain what they mean by this vague term. Each focuses on a specific part of the hierarchical architecture of PP and on a particular set of properties of such posits and decisively arrives at a final verdict: that PP is nonrepresentational or that 4E cognition is representational. That is, both present instances of (non) representationality arguments explicated in section 3.

Let us first contemplate the argument made in (Hohwy 2018). The argument schema presented is as follows:

- 1) PP is necessarily representational.
- 2) 4E cognition, correctly understood, is representational because PP, given its representational nature, can “encompass phenomena highlighted in debates on 4E cognition” (Hohwy 2018, 130).

⁵ Kirchhoff (2018) additionally argues that the PP framework is compatible with the constitutive, cognitive-affective inseparability, and metaplasticity theses of what he thinks 4E cognition is, and Hohwy (2018) also argues for the inferentiality of 4E cognition. In this section, considering the scope of the present study, I shall only focus on their representationality arguments.

3) Therefore, 4E cognition and PP are compatible.

To argue for the first premise, he demonstrates how representations are, as it were, *sine qua non*s for the brain's PEM strategy. He does not explicitly specify any cognitive domain or any parts of the hierarchical architecture and appears to deal with the whole multilayered structure. Representational posits of PP play the functional role of mirroring nested causal complexities of the external world. The better the external world is represented, the more accurate predictions generated by the model will be, making such representations essential for PP. For the second premise, he points out that due to the FEP (see Friston 2010; Friston and Stephan 2007), the surprise is relative to the model formed directly by the organism's phenotype. This endows bodily states and the environment with central roles in cognition. The content is then organism-salient, and the body and natural environment of the organism are not trivial but play fundamental roles. The content of those posits speaks to the tenets of 4E cognition. Thus, since he takes representations for granted from the first premise, he declares that 4E cognition is representational depending on the FEP, which renders PP compatible with essential tenets of 4E cognition. He takes the representational function of the generative model and draws a conclusion about representational content congenial to 4E cognition.

The first concern as to the validity of his argument stems from what Hohwy means by the term "representation," which he argues is necessary for PP. Hohwy mentions two hallmarks of representations: action-guidance and detachability (Hohwy 2018, 135). He derives the functional property of action-guidance from active inference, but he does not elaborate more on in what sense those representational posits are 'detached' or even on what it means for a posit to be 'detached.'⁶ It is highly controversial whether these two properties turn any posits into representations. When arguing that PP is representational, Gładziejewski (2015, 12) claims that 'acting-guiding'

⁶ This point is more crucial than it seems because in the literature one can find various understandings of "detachment" and "decoupleability," which might be decisive in understanding what kinds of representations Hohwy refers to. For the present study, it is enough to raise this question and not to go into detail to avoid further discussion. See Rowlands (2012) and Gallagher (2017, 88-103) for more discussion on the significance of these terms in the representation wars.

and ‘detachability’ are not jointly sufficient to be counted as representation. He seems to refer to a pretheoretic understanding of internal representations (i.e., mirroring nature) and does not go into theoretical details to prove that these are robust representations according to any given definition.

Next, even if one assumes that his premises are correct, one must raise the question of which of the 4E notions Hohwy’s framework is compatible with. Hohwy acknowledges that he refers to “somehow deflated 4E notions” (Hohwy 2018, 130). However, details of these “somehow deflated 4E notions” remained obscure. In a paper of this scope, it is not feasible to attempt to review all versions of 4E cognition and to find the ones Hohwy refers to. Nevertheless, certain examples can be analyzed to illustrate the possible positions Hohwy might *not* have taken. Then, first, it is clear that for Hohwy these posits are content-bearing states because he already argued that the content of a posit is organism-salient. Thus, his proposal does not carry any kinship to Hutto and Myin (2013) because for them, a great many number of cognitive phenomena are contentless. His account also stands in strong opposition to those for whom just a covariance relation⁷ must suffice for at least some cognitive tasks (especially lower-level ones) such as Gallagher (2017), Chemero, (2009), Noë (2004), and Van Gelder (1995). Though action-guiding, Hohwy’s posits have the function of building a rich and fragile representation of the external world, which, without further ado, situates him against a variety of enactivists such as Varela et al. (1991).

It is somehow painless to point out whose representational posits are not compatible with Hohwy’s. However, if his argument is not that PP is compatible with what I call 4E cognition, what he refers to by “somehow deflated 4E cognition” must be found in the literature. This seems, however, slightly more challenging. He names in the article in question some 4E cognition proponents such as Varela (1991), Gallagher (2005), and Thompson (2007) with whom his framework might “have contact” and others such as

⁷ Note that while the main tendency among contenders of debates on representations is to think that a covariance relation is not enough to provide natural content for representational posits, some can still argue that it is. Thus, to avoid further complications and discussions resulting from these different understandings of philosophers, I refer to those who use the term “contentless” and to those who prefer to use the term “just covariance” separately.

Van Gelder (1995) to whom his theory “speaks.”⁸ Above I have already shown that his representational posits are incompatible with those in these accounts. Possible candidates would be among those 4E notions deploying representational terminology explicitly such as Clark (2008), Wheeler (2005), Goldman (2012), Goldman and de Vignemont (2009), Bickhard (2000), Barsalou (1999), Mandik (2005), Prinz (2009), and Hutchins (1995). However, one must be careful in claiming that (Hohwy 2018)’s account is compatible with these accounts based on the fact that both use explicit representational terminology because, for example, Wheeler (2005; 2008) does away with decoupleability as a criterion for minimal representation. If Hohwy’s ‘detachment’ means ‘decoupleability’ in the sense that a representational posit can also perform its function of guiding intelligent behavior (or action-guiding as stated in (Hohwy 2018)) even when the represented extracranial target is absent, then this jeopardizes their compatibility at least for some lower level cognitive phenomena. That is, Wheeler can explain certain cognitive phenomena with posits that are not decoupleable with their target, but Hohwy (2018) cannot. Furthermore, since it is clear that Hohwy’s posits carry extracranial bodily *content*, this might be compatible with Mandik (2005), but whether these posits have the bodily *format* (not only content)⁹ required by Goldman (2012), Goldman and de Vignemont (2009), Barsalou (1999), etc. is not palpable in (Hohwy 2018).¹⁰

⁸ He also mentions (Clark 1997) and (Aydede and Robbins eds. 2009). However, the central arguments in these works concern intracranialism, not representationalism. Since the present study is exclusive to the representationality discussion, both works are intentionally excluded.

⁹ For a detailed discussion of and substantial criticism concerning “bodily content” and “bodily format,” see Hutto (2013) and Goldman and de Vignemont (2009), and for a more general discussion of bodily representations and their controversial status in the 4E tradition, see Alsmith and de Vignemont (2012).

¹⁰ However, while in (Hohwy 2018) it is hardly possible to understand whether these representational posits are of bodily formats, other texts on PP in the literature reveal that such posits can have not only bodily content but also bodily formats. For example, Gładziejewski wrote that “generative models [...] could bring about patterns of neural activity that resemble those that accompany perception and action” (2017, 106) when performing certain detached cognitive tasks such as imagining, counterfactual reasoning, dreaming, etc.

We are thus only left with a limited number of weakened versions of 4E cognition. If he bites the bullet and reduces his claim to cover only some highly deflated versions of 4E cognition or maybe even only the version of 4E cognition described in (Hohwy 2018), then he faces the risk of trivializing his position in this debate, because what is the point of claiming that versions of 4E cognition that accept representations are representational? He might also be trivializing 4E cognition in general, because I see no reason why an embodied cognitive scientist would accept those overtly representational accounts *qua* accounts of 4E cognition.¹¹

Please note, however, that the point of the argument is to demonstrate how insufficient arguments for the compatibility of (conservative) PP and 4E cognition are if they depend on representational 4E cognition and not that Hohwy's framework is only compatible with certain deflated notions of 4E cognition. This is already rightly stated by Schlicht in his critical note to (Hohwy 2018): "PEM can only be compatible with the moderate 4E approaches" (Schlicht 2018, 219).

Let us now turn to Kirchhoff's argument that PP is compatible with 4E cognition because PP can embrace the nonrepresentational thesis of 4E cognition. The nonrepresentational thesis claims that "[t]he sensorimotor profile of organisms is sufficient for at least some kinds of cognitive activities, thus replacing the need for organisms to construct complex internal mental representations of the outside environment" (Kirchhoff 2018, 244). As the saying goes, one man's *modus ponens* is another's *modus tollens*. Kirchhoff's argument touches on the same point as Hohwy's but from a different perspective. His first premise points out active inference and the FEM context, which emphasizes the continuous coupling of the body, environment, and agent in the PP framework. This idea is a familiar one since this dynamical coupling represents one of the initial ideas of 4E cognition and is supported by almost any defender of 4E cognition (see, e.g., Chemero 2009; Gallagher

¹¹ Some of the theories' positions in the 4E cognition tradition are highly controversial because they explicitly make use of robust representations in their explanations. For instance, Walter renders many of such accounts such as Goldman (2012) as "not embodied at all" (Walter 2014, 246) while Alsmith and de Vignemont consider them "weakly embodied" (2012, 3).

2017). By virtue of active inference, for Kirchhoff, the body and environment become not only coupled but also one interacting system. Thus, the cognitive function of the mind endowed with PP is not to think, but rather to act and engage with the world. With this initial support from active inference, he derives another form of support from surprises relative to a model, which leads him to the same conclusion as Friston that “an agent does not *have* a model of its world— it *is* a model” (Friston 2013, 213, emphasis in original). This reading of a model is more inclusive, loose, and pretheoretic; it does not imply strict separation. These peculiarities of PP are compatible with 4E cognition.

His second premise is directed against the argument that to perform PEM, which is also used to minimize nested causal complexity, there must be representations due to self-evidencing.¹² Hohwy (2014; 2017) thinks that this leads to the idea of a brain that is secluded from its environment and extracranial body. Kirchhoff admits that there is self-evidencing here too, but synergy, according to (Kirchhoff 2018), provides this connection without appealing to representations. He offers the idea of ‘synergy’ (see Kelso 1995; 2009), thanks to which in a nontrivial way, “organisms can minimize complexity” (Kirchhoff 2018, 250). ‘Synergy,’ he proposes, is “*shortly-lived assembly*” (Kirchhoff 2018, 250, emphasis added). Synergies are “not as static representational as motor programs” (Riley et al. 2012, 23). Assuming

¹² Hohwy (2014; 2017) notices that in the PP framework, when one has a hypothesis that explains some of its evidence, it also provides evidence for itself. In his own words:

The internal model that generates hypotheses that over time makes the evidence most likely, and does so most precisely and simply, will have its own evidence maximized. That is, as a model generates hypotheses that explain away occurring surprising evidence (i.e., minimize prediction error) it maximizes the evidence for itself. Prediction error minimization thus constitutes self-evidencing. This is then the doctrine of the self-evidencing brain (Hohwy 2014, 6).

To Hohwy, “this enforces an evidentiary boundary between it and the external causes of sensory input harbored in the environment and in the rest of the body” (Hohwy 2014, 1). This leads to a strict separation between the brain and its environment including the extracranial body, which contradicts the underlying claims of 4E cognition. Thus, self-evidencing in the PP framework creates the need for representations to bridge the gap between these elements.

that this is true, there is no gap between the world, the body, and the agent, which obviates the need for representations to bridge the gap between them. Thus, we need not have representations for the PP hypothesis to be true, and PP is perfectly compatible with 4E cognition. In arguing this, what he has in mind is obviously low-level models. Low-level models are thought to be responsible for short term, on-the-fly, and almost reflexive activities. Even the passage he quotes from Clark reveals this fact: “... model’ means in *low-level* free energy minimization accounts...” (Clark 2016, 14, emphasis added).

In considering these points, he argues against the functional property “decoupleability” of those posits. The representational posits of PP (as I argue, only the ones of low-level models in PP) are not decoupleable since they are in a causal coupling with external states of affairs. These representational posits do not bear any content and function as a merely causal mediator “because internal and external states cause one another in a circular and reciprocal fashion” (Kirchhoff 2018, 251). “They do not seem to imply the presence of inner model or content-bearing states” (Clark 2016, 14). Kirchhoff’s position could point to another interesting conclusion. He seems to understand representations in a more theoretical sense because he does not directly reject the existence of representational posits. He sets out to illustrate how they are causally coupled with external states, and “decoupleability” seems to be his criterion for a representation,¹³ but this is not explicitly stated in (Kirchhoff 2018).

Another point undermining Kirchhoff’s nonrepresentationality argument relates to the fact that while for Hohwy’s argument, pointing out a single representation proves that cognition involves representations, for Kirchhoff, it does not. He must have shown that not only the generative model but also all other posits of PP are not representations. However, he only discusses the generative model while remaining silent about the other posits listed above, which may be representations. To him, if any of these posits happens to be decoupleable, this invalidates his argument.

Finally, we can ask which version of 4E cognition he refers to. If he argues that certain cognitive phenomena are nonrepresentational and thus

¹³ Remember, for instance, that Wheeler (2005; 2008) contends that minimal representations do not need to exhibit ‘decoupleability.’

contentless, then he loses sight of a myriad of nonradical versions of 4E cognition. Indeed, only some extremely radical versions of enactivism and embodiment deploy contentless posits such as Hutto and Myin's (2013). Kirchhoff's nonrepresentationality thesis claims that "*at least some kinds of cognitive activities*" (2018, 244, emphasis added) are nonrepresentational and thus contentless. However, proponents of radical 4E cognition make use of contentless posits in explaining cognitive phenomena of various levels. Thus, unless Kirchhoff specifies what "*at least some kinds of cognitive activities*" are and to what extent he can explain cognitive phenomena with contentless PP architecture, it also appears impossible to determine exactly which versions of 4E cognition Kirchhoff refers to.

In sum, Hohwy's posits have the representational function of mirroring nature, and their content is organism-salient. Since he takes representations from this representational function, he does not seem to bother contemplating whether they are content-bearing states or not. Starting from the cognitive function of a mind employing PP, Kirchhoff derives the nondecouplability of representational posits: they are causal mediators. His second argument relates to the so-called content of these posits. For Kirchhoff, surprise being relative to a given model means that an organism does not harbor a model of the environment, but becomes the model of its environment. Accordingly, such posits do not bear any content. Moreover, while Kirchhoff seems to only discuss low-level models, Hohwy deals with the whole hierarchical architecture performing PEM. Finally, in both articles, it is not palpable which versions of 4E cognition are being considered. Possible candidates for Hohwy's account might only be searched among certain deflated versions, and Kirchhoff's account can be compatible to radical versions of 4E cognition.

5. Conclusion

In section 2, a selective and brief overview of PP and 4E cognition is provided, and how the question of their relation is tied to the features of representational posits they deploy is demonstrated. In section 3, the general schema of the (non) representationality argument mostly appealed to in arguing for the compatibility of PP and 4E cognition is given, and it is

shown why any instance of such an argument falls short of proving that PP and 4E cognition are compatible. In the last section, Hohwy and Kirchhoff's arguments are analyzed in detail to illustrate how the criticism raised here is a significant one and the consequences of ignoring it.

The argument presented here is a good one only insofar as it can show that if it depends on either nonrepresentational PP or representational 4E cognition, any argument for the compatibility of PP and 4E cognition can become much more complicated than treated so far in the literature of cognitive science and philosophy of mind. If it is a good one, its significance does not lie in detecting exactly which version of 4E cognition or which kinds of representational posits the authors in question refer to, but in genuinely demonstrating why such arguments are always insufficient in what they argue for.

It is definite that the representational posits of PP are radically different from the older ones in various respects and that the anti-representationalism of 4E cognition is not homogenous. Given this, arguments for nonrepresentational PP or representational 4E cognition fail to appreciate diversity and heterogeneity in this pertinent debate. Rather, an ontological analysis of these representational posits and of their capacities to satisfy skepticisms of 4E cognition should be central if one aims to argue for the compatibility of 4E cognition and PP.

To conclude, neither of the theories have a uniform understanding of the debate. When arguing against representations, proponents of 4E cognition have different and distinct ideas of representations in mind. When evaluating PP, its proponents refer to various parts of the whole multi-layered structure. This variety hinders the validity of any argument for the compatibility of these theories if they propose instances of the (non) representationality debate sketched in section 3. Hence, without depending on any definition of representation, one must refer to similarities and dissimilarities in the properties of these posits in both theories. With this paper I hope to have shown that the growing tendency in the literature to argue about the existence of representations in these theories to establish their compatibility should be abandoned because it seems quite impossible to resolve the complexities of representationalism debates on PP and 4E cognition in a single argument. For this reason, this paper contends that such

arguments will always be partial and fail to be conclusive. Instead, an effort to reveal the representational properties of the posits of each particular theory and to compare these properties shall better serve this debate.

References

- Adams, Fred, and Aizawa, Ken. 2008. *The Bounds of Cognition*. Oxford: Blackwell. <https://doi.org/10.1002/9781444391718>
- Agre, Philip, and Chapman, David. 1987. "Pengi: An Implementation of a Theory of Activity." *Proceedings of the Sixth National Conference on Artificial Intelligence* 1: 268–72.
- Alsmith, Adrian J. T., and de Vignemont, Frederique. 2012. "Embodying the Mind and Representing the Body." *Review of Philosophy and Psychology* 3(1): 1–13. <https://doi.org/10.1007/s13164-012-0085-4>
- Aydede, Murat, and Robbins, Philip eds. 2009. *The Cambridge Handbook of Situated Cognition*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511816826>
- Barsalou, Lawson W. 1999. "Perceptual Symbol Systems." *Behavioral and Brain Sciences* 22(4): 577–609. <https://doi.org/10.1017/s0140525x99002149>
- Barsalou, Lawson W. 2008. "Grounded Cognition." *Annual Review Psychology* 59(1): 617–45. <https://doi.org/10.1146/annurev.psych.59.103006.093639>
- Bickhard, Mark H. 2000. "Information and Representation in Autonomous Agents." *Cognitive Systems Research* 1(2): 65–75. [https://doi.org/10.1016/s1389-0417\(99\)00007-8](https://doi.org/10.1016/s1389-0417(99)00007-8)
- Brooks, Rodney A. 1991. "Intelligence Without Representation." *Artificial Intelligence* 47(1–3): 139–59. [https://doi.org/10.1016/0004-3702\(91\)90053-m](https://doi.org/10.1016/0004-3702(91)90053-m)
- Chemero, Anthony. 2009. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/8367.001.0001>
- Clark, Andy. 1997. *Being There*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/1552.001.0001>
- Clark, Andy. 2008. *Supersizing the Mind*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195333213.001.0001>
- Clark, Andy. 2013. "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences* 36(3): 181–253. <https://doi.org/10.1017/s0140525x12000477>
- Clark, Andy. 2015a. "Predicting Peace: The End of the Representation Wars-A Reply to Michael Madary." In *Philosophy and Predictive Processing*, edited by

- T. Metzinger and W. Wiese. 10. Frankfurt am Main: MIND Group.
<https://doi.org/10.15502/9783958570979>
- Clark, Andy. 2015b. "Radical Predictive Processing." *The Southern Journal of Philosophy* 53: 3–27. <https://doi.org/10.1111/sjp.12120>
- Clark, Andy. 2016. *Surfing Uncertainty*. Oxford: Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780190217013.001.0001>
- Dolega, Krzysztof. 2017. "Moderate Predictive Processing." In *Philosophy and Predictive Processing*, edited by T. Metzinger and W. Wiese. 10. Frankfurt am Main: MIND Group. <https://doi.org/10.15502/9783958573116>
- Dretske, Fred. 1988. *Explaining Behavior*. Cambridge, MA: MIT Press.
- Fodor, Jerry. 1975. *The Language of Thought*. Cambridge: MIT Press.
- Foglia, Lucia, and Grush, Rick. 2011. "The Limitations of a Purely Enactive (Non-Representational) Account of Imagery." *Journal of Consciousness Studies*, 18 (5-6): 35–43.
- Friston, Karl, and Stephan, Klaas. 2007. "Free-Energy and the Brain." *Synthese* 159: 417–58. <https://doi.org/10.1007/s11229-007-9237-y>
- Friston, Karl. 2010. "The Free-Energy Principle: a Unified Brain Theory?" *Nature Reviews. Neuroscience* 11(2): 127–38. <https://doi.org/10.1038/nrn2787>
- Friston, Karl. 2013. "Life as We Know It." *Journal of The Royal Society Interface* 10: 20130475. <https://doi.org/10.1098/rsif.2013.0475>
- Gallagher, Shaun. 2005. *How the Body Shapes the Mind*. Oxford: Oxford University Press. <https://doi.org/10.1093/0199271941.001.0001>
- Gallagher, Shaun. 2008. "Are Minimal Representations Still Representations?" *International Journal of Philosophical Studies* 16: 351–69.
<https://doi.org/10.1080/09672550802113243>
- Gallagher, Shaun. 2017. *Enactivist Interventions: Rethinking the Mind*. Oxford: Oxford University Press.
<https://doi.org/10.1093/oso/9780198794325.001.0001>
- Gładziejewski, Paweł. 2015. "Predictive Coding and Representationalism." *Synthese* 193(2): 559–82. <https://doi.org/10.1007/s11229-015-0762-9>
- Gładziejewski, Paweł. 2017. "Just How Conservative is Conservative Predictive Processing?" *Hybris. Revista de Filosofia* 38: 98–122.
- Goldman, Alvin, and de Vignemont, Frederique. 2009. "Is social cognition embodied?" *Trends in Cognitive Sciences* 13: 154–59.
<https://doi.org/10.1016/j.tics.2009.01.007>
- Goldman, Alvin. 2012. "A moderate approach to embodied cognitive science." *Review of Philosophy and Psychology* 3: 71–88. <https://doi.org/10.1007/s13164-012-0089-0>
- Grush, Rick. 1997. "The architecture of representation." *Philosophical Psychology* 10: 5–24. <https://doi.org/10.1080/09515089708573201>

- Grush, Rick. 2004. "The emulation theory of representation: Motor control, imagery, and perception." *Behavioral and Brain Sciences* 27: 377–442.
<https://doi.org/10.1017/s0140525x04000093>
- Haselager, P., De Groot, Andre, and Van Rappard, Hans. 2003. "Representation-alism vs. anti-representationalism: a debate for the sake of appearance." *Philosophical psychology* 16(1): 5–24.
<https://doi.org/10.1080/0951508032000067761>
- Hohwy, Jacob. 2013. *The Predictive Mind*. Oxford: Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199682737.001.0001>
- Hohwy, Jacob. 2014. "The Self-Evidencing Brain." *Noûs* 50(2): 1–27.
<https://doi.org/10.1111/nous.12062>
- Hohwy, Jacob. 2017. "How to Entrain Your Evil Demon." In *Philosophy and Predictive Processing*, edited by T. Metzinger and W. Wiese. 10. Frankfurt am Main: MIND Group. <https://doi.org/10.15502/9783958573048>
- Hohwy, Jacob. 2018. "The Predictive Process Hypothesis." In *The Oxford Handbook of 4E Cognition*, edited by A. Newen, L. de Bruin and S. Gallagher, 129–45. Oxford: OUP.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.
- Hutto, Daniel, and Myin, Eric. 2013. *Radicalizing Enactivism: Basic Minds without Content*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/9780262018548.001.0001>
- Hutto, Daniel, and Myin, Eric. 2018. "Going Radical." In *The Oxford Handbook of 4E Cognition*, edited by A Newen, L de Bruin, and S. Gallagher, 95–115. Oxford: OUP.
- Hutto, Daniel. 2013. "Exorcising Action Oriented Representations: Ridding Cognitive Science of its Nazgûl." *Adaptive Behavior* 21(3): 142–50.
<https://doi.org/10.1177/1059712313482684>
- Kelso, Scott. 1995. *Dynamic Patterns*. Cambridge, MA: MIT Press.
- Kelso, Scott. 2009. "Synergies: Atoms of Brain and Behavior." In *Progress in Motor Control: A Multidisciplinary Perspective*, edited by D. Sternad, 83–91. New York: Springer.
- Kiefer, Alex, and Hohwy, Jacob. 2017. "Content and Misrepresentation in Hierarchical Generative Models." *Synthese* 195: 2387–415.
<https://doi.org/10.1007/s11229-017-1435-7>
- Kirchhoff, Michael D. 2018. "The Body in Action: Predictive Processing and the Embodiment Thesis." In *The Oxford Handbook of 4E Cognition*, edited by A Newen, L de Bruin, and S. Gallagher, 243–60. Oxford: OUP.
- Kirchhoff, Michael D., and Robertson, Ian. 2018. "Enactivism and Predictive Processing: a Non-Representational View." *Philosophical Explorations* 21(2): 264–81. <https://doi.org/10.1080/13869795.2018.1477983>

- Lakoff, Georg, and Johnson, Marl. 1980. *Metaphors We Live By*. Chicago, IL: University of Chicago.
- Lakoff, Georg, and Johnson, Marl. 1999. *Philosophy in the Flesh*. NY: Basic Books.
- Madary, Michael. 2015. "Extending the Explanandum for Predictive Processing—A Commentary on Andy Clark." In *Philosophy and Predictive Processing*, edited by T. Metzinger and W. Wiese. 10. Frankfurt am Main: MIND Group. <https://doi.org/10.15502/97839558570313>
- Mandik, Pete. 2005. "Action-Oriented Representation." In *Cognition and the Brain: The philosophy and Neuroscience Movement*, edited by Brook, A., and Akins, Kathleen, 284–305. Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511610608.009>
- Millikan, Ruth. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, Mass.: MIT Press.
- Millikan, Ruth. 1993. *White Queen Psychology and Other Essays for Alice*. Cambridge, Mass.: MIT Press.
- Millikan, Ruth. 1995. "Pushmi-Pullyu Representations." *Philosophical Perspectives* 9: 185–200. <https://doi.org/10.2307/2214217>
- Newen, Albert, de Bruin, Leo, and Gallagher, Shaun, eds. 2018. *The Oxford Handbook of 4E cognition*. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780198735410.001.0001>
- Noë, Alva. 2004. *Action in Perception*. Cambridge: MIT Press.
- Orlandi, Nico. 2014. *The Innocent Eye: Why Vision is not a Cognitive Process*. Oxford, NY: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199375035.001.0001>
- Prinz, Jesse. 2009. "Is Consciousness Embodied?" In *The Cambridge Handbook of Situated Cognition*, edited by Aydede, Murat and Robbins, Philip, 419–37. Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511816826.022>
- Pylyshyn, Zenon W. 1984. *Computation and Cognition*. Cambridge, MA: MIT press.
- Ramsey, William. 2007. *Representation Reconsidered*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511597954>
- Riley, Michael A, Schockley, Kevin, and Van Orden, Guy. 2012. "Learning from the Body about the Mind." *Topics in Cognitive Science* 4: 21–34. <https://doi.org/10.1111/j.1756-8765.2011.01163.x>
- Rowlands, Mark. 2006. *Body Language*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/1643.001.0001>
- Rowlands, Mark. 2012. "Representing Without Representations." *AVANT* 3(1): 133–44.

- Schlicht, Tobias. 2018. "Critical Note: Cognitive Systems and the Dynamics of Representing-in-the-World." In *The Oxford Handbook of 4E Cognition*, edited by A Newen, L de Bruin, and S. Gallagher, 321–32. Oxford: OUP.
- Svensson, Henrik, and Ziemke, Tom. 2005. "Embodied Representation: What are the Issues?" In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, edited by B G Bara, L Barsalou, and M Bucciarelli, 2116–21. Stresa: Italy.
- Thelen, Esther, and Linda, Smith. 1994. *A Dynamic Systems Approach to the Development of Cognition and Action*. Cambridge: MIT Press.
<https://doi.org/10.7551/mitpress/2524.001.0001>
- Thompson, Evan. 2007. *Mind in Life: Biology, Phenomenology, and the Science of Mind*. Harvard: Harvard University Press.
- Thompson, Evan. 2011. "Replies to Commentaries." *Journal of Consciousness Studies* 18 (5-6): 76–223.
- Van Gelder, Tim. 1995. "What Might Cognition Be, If Not Computation?" *Journal of Philosophy* 92(7): 345–81. <https://doi.org/10.2307/2941061>
- Varela, F., Thompson, Evan, and Rosch, Eleanor. 1991. *The Embodied Mind*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/6730.001.0001>
- Walter, Sven. 2014. "Situated Cognition: A Field Guide to Some Open Conceptual and Ontological Issues." *Review of Philosophy and Psychology* 5(2): 241–63.
<https://doi.org/10.1007/s13164-013-0167-y>
- Wheeler, Michael. 2005. *Reconstructing the Cognitive World: The Next Step*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/5824.001.0001>
- Wheeler, Michael. 2008. "Minimal Representing: A Response to Gallagher." *International Journal of Philosophical Studies* 16(3): 371–76.
<https://doi.org/10.1080/09672550802113276>

A Novel Reading of Thomas Nagel’s “Challenge” to Physicalism

Serdal Tmkaya*


Received: 12 June 2020 / Accepted: 18 November 2020

Abstract: In passing remarks, some commentators have noted that for Nagel, physicalism is true. It has even been argued that Nagel seeks to find the best path to follow to achieve future physicalism. I advance these observations by adding that for Nagel, we should discuss the consciousness problem not in terms of physical and mental issues but in terms of our desire to include consciousness in an objective/scientific account, and we can achieve this only by revising our self-conception, i.e., folk psychology, to develop a more detached view of experience. Through the project of objective phenomenology, Nagel aims to achieve some sort of objective, detached, and scientific explanation of the subjective nature of experience. This project seeks to make the truth of physicalism intelligible and consciousness more amenable to scientific study, potentially raising an even broader concept than the one physicalism originally proposes.

Keywords: Folk psychology; Nagel; objective phenomenology; physicalism; science of consciousness.

* Middle East Technical University and Ardahan University

 <https://orcid.org/0000-0002-2453-4184>

 Ardahan University Yenisey Kamps, amlıatak, Ardahan Department of Philosophy, 75002 Ardahan, Turkey

 serdal82@gmail.com

1. Introduction

“[P]hysicalism ... repels me although I am persuaded
of its truth.” (Nagel 1965, 356)

Thomas Nagel has been claimed (i) to argue against physicalism (e.g., Thomas 2009; Sundström 2002), (ii) to have changed his position from physicalism to nonphysicalism and later to anti-physicalism from 1965 to 1998 (e.g., Nagasawa 2003, 377; cf. Foss 1993, 726) or from 1986 to 2012 (Seager 2014, 10, n12), and (iii) to deny the possibility of giving an objective account of consciousness (e.g., Dennett 1991, 71; Bond 2005, 129–30; McHenry and Shields 2016, 497).¹ Some commentators have noted that Nagel does not in fact claim that physicalism is false (D’Oro 2007, 170). For example, Seager asserts that

Nagel is officially agnostic about the truth of physicalism, or even leans towards accepting it, but takes it for granted that absent a plausible route towards establishing reductive epistemological dependence, arguments in favour of a physicalist solution to the mind-body problem are just ‘sidestepping it’. (Seager 2014, 10)

In an associated note (n12) on the same page, he adds that “At least, that was true at the time Nagel wrote the famous bat paper and in Nagel (1986); he seems to have definitively rejected physicalism in his latest work, Nagel (2012).” As he says, it is also possible to think of Nagel as leaning toward accepting physicalism.

On rare occasions, it has been acknowledged that he tends to believe that physicalism is true but is still suspicious of its sufficiency. Tim Crane observed that “Nagel’s view was that physicalism is true, but that we cannot fully understand it” (2007, 23). More importantly, Stubenberg argues that Nagel wants to clear the path for a future physicalism (see his 1998). I further their observation by adding that for Nagel we should discuss the consciousness problem not in physical and mental terms but from our desire to include consciousness in an objective/scientific account and that we can achieve this only by revising our self-conception, i.e., folk psychology, to develop a more detached view of experience.

¹ See section 3 for many other references.

Through the project of objective phenomenology, Nagel aims to achieve some sort of objective, detached, and scientific explanation of the subjective aspect of experience (1986, chaps. 1–2, see also 1974, 449). The project intends to make the truth of physicalism intelligible and consciousness more amenable to scientific study.

This article places Thomas Nagel's ideas regarding physicalism in the context of his proposal for objective phenomenology. In fact, a detailed discussion of the objective phenomenology project provides the basis for Nagel's lesser known critique of folk psychology. Unfortunately, a large majority of the literature on the philosophy of mind focused on Nagel is silent with respect to his austere criticisms of the deficiencies of our self-conception.² Once we understand this triangular relationship, it is much easier to understand that embracing physicalism is quite compatible with Nagel's general framework. Though Nagel has said that "consciousness is what makes the mind–body problem really intractable" (1974, 435), the objective phenomenology project is meant to make it tractable within a physicalist framework. It is true that consciousness is obstinate to an objective characterization, but it is not impossible to remove its resistance if we sufficiently revise our current framework of folk psychology (Nagel 1993, 2002). Let me start by providing initial definitions for consciousness and folk psychology before discussing the relationships between them.

Whatever else consciousness is, it is typically presented as something unitary, accessible to the privileged first-person view, intentional, and qualitative. Folk psychology is, roughly, the framework underlying generalizations made by lay people to understand and predict the behaviors of other people and their own behaviors. It also reflects the familiar conception of mind that ordinary people endorse, and this conception infiltrates the usual language of philosophers of mind when they talk about consciousness. Through this envisaged revision in folk psychology, physicalism's seeming wanting disappears, as our standards to judge the soundness of any physicalist theory will change. Thus, physicalism lives up to our expectations. The naïve preconception of the mind and body is restructured. The

² Nagel uses the terms "folk psychology," "our self-conception," "standard mentalistic idioms," and "traditional conception of mind" interchangeably. This is not a problem of content, but it might be confusing for readers.

ossified conception of the mind fails to adapt to new knowledge that we have gained about the deficiencies of our self-conception and is likely to be significantly updated and may die out (see Nagel 1998). We shall discuss these points in the following sections.

2. Nagel embraces physicalism

"I am inclined to believe that some weak physicalist theory of the third type is true ..." (1965, p. 340)

Physicalism is one of the most protean terms developed in recent philosophy. There are thousands of philosophers who are physicalist from one point of view, but anti-physicalist from another. For some, physicalism suggests that phenomenal consciousness is an illusion. More accurately, "if physicalism is true, phenomenal consciousness must be an illusion" (Tartaglia 2016, 236). For others, we can and should be realists about our experiences but also physicalists of some sort. For instance, Nagel describes Galen Strawson as a materialist of an anti-reductionist type and a realist about experience:

However, Strawson is a materialist and does not think that your self could exist apart from your central nervous system. He holds that your experiences are events in your brain, and that if there is a self which is their subject it too must be in the brain. But he is a materialist of an unusual kind: a realist about experience and an anti-reductionist. (Nagel 2009)

Strawson is a realist about experience and an anti-reductionist on the mind-body problem as Nagel is. Despite this, Nagel says that Strawson is an unusual kind of materialist. According to this line of reasoning, Nagel should call himself a physicalist. This is not unexpected, since contrary to what so many philosophers believe, Thomas Nagel tends to believe that a *weaker* form of physicalism is *true* (1965, 340 and 356). Strong physicalism represents a type identity theory for him. The right alternative is some sort of token physicalism. This reflects a strong version of token materialism and is occasionally referred to as neutral monism or dual-aspect theory (or

even pansychism once upon a time) by him (2012, 4–5; cf. Pernu 2017, 6).³ For him, the mental is strictly dependent on the physical:

The mind-brain case seems a natural candidate for such treatment because what happens in consciousness is pretty clearly supervenient on what happens physically in the brain. In the present state of our conceptions of consciousness and neurophysiology, this strict dependence is a brute fact and completely mysterious. (Nagel 2002, 207)

Nagel says that “Materialism is the currently dominant form of reductionism, and it reduces the mental to the physical via the reduction of the mental to the biological” (Nagel et al. 2016, 394). When he uses the word materialism as shorthand for “reductive physicalism” or for specific forms of “naturalism,” he rejects it. He dismisses them on the grounds that they leave something important about consciousness unexplained. He sometimes claims that mentalistic concepts are indispensable (see, e.g., 2016, 400) and at other times seriously entertains the idea that all of our mentalistic concepts with all the principles of our self-conception will not survive the next century *intact* (1998).

This brings us to a difficult question: is physicalism true for Nagel or not? He sometimes argues that mentalistic language is indispensable but at other times contends that physicalism must be true. Some philosophers see an equivocation between the two notions of physicalism illustrated in Nagel’s publications, as Torin Alter notes (but mistakenly rejects):

There are other ways to interpret Nagel’s view in WLBB [1974] about the status of physicalism. For example, one possibility is that he equivocates on “physicalism”: in some places he uses the term to refer to certain reductionist theories that he outright rejects, and in others he uses it to refer to physicalist theories that he believes are compatible with S and possibly true.⁴ (Alter 2002, 155, n11)

³ McGinn noted that “Something close to anomalous monism is tentatively endorsed by Thomas Nagel in ‘Physicalism’” (1980, 202, n3).

⁴ S denotes the general principle that “Experiences are subjective: understanding their true nature requires having or imaginatively adopting the viewpoint of the experiencing creature” (2002, 147).

By blurring the distinctions between materialism as such, reductive physicalism, and naturalism, Nagel causes considerable confusion for his readers. We must first remedy this confusion. We will start with physicalism, continue with objectivity, and end with folk psychology. My argument, as a first approximation, can be summarized as follows. Due to conspicuous deficiencies of folk psychology, i.e., our self-conception, it is difficult for us to imagine that mental states are physical states and that physicalism is true. The objective phenomenology project offers the capacity to make this truth intelligible.

Nagel tries to teach us to think of physicalism problem in terms of the objective–subjective relationship. This relationship is not a polarized one but shows gradation. Moreover, the two sides of the relation are parts of a continuum (see section 3). We have evidence that physicalism is true, but we do not know how and why it is true: "... I think we also have some reason to believe that sensations are physical processes, without being in a position to understand how" (1974, 448). We do not fully know its nature. What he argues against are *certain* sorts of physicalism as follows: scientific, reductionist, and functionalist ones (Nagel 2002). He levels charge against these versions of physicalism; an approach based upon common sense, assuming the possibility of logical reductionism, granting the correctness of our self-conception, rather than explaining subjective aspects of experience ignoring it (Nagel 1965, 1970).

The point for Nagel, as stated above, is not physicalism as such but the objectivity problem:

We cannot genuinely understand the hypothesis that their nature is captured in a physical description unless we understand the more fundamental idea that they have an objective nature (or that objective processes can have a subjective nature). (1974, 448) (*Italics original*)

In the debate over consciousness, Nagel should be regarded as a nonsubjectivist: "My aim is to clarify and explore this question and to try, for certain domains of thought, to defend what I shall call a rationalist answer against what I shall call a subjectivist one" (1997, 3). Nagel stipulates that if something is physical, then it must be objective (1974, 449, fn 15). In addition, he, in his 1965 article entitled "Physicalism," explicitly states that

a weaker form of physicalism is likely to be true. However, it would be better, he argues in “What is It Like to Be a Bat?” (1974) and “Conceiving the Impossible and the Mind-Body Problem” (1998), to conceptually revise our mentalistic ideas. Hence, we should see him as a *revisionary* materialist (cf. Allen-Hermanson 2015, 59–63; cf. Bickle 1992, 1998, chap. 6; cf. Sundström 2018).

Let us proceed to see what the problem is with folk psychology. Below is Nagel’s surprisingly harsh criticism of it based on the fact that our mentalistic ideas have naturally evolved through nonscientific functions. He then goes on saying that:

Our dealings with and declarations to one another require a *specialized* vocabulary, and although it serves us moderately well in *ordinary* life, its *narrowness* and *inadequacy* as a psychological theory become evident when we attempt to apply it in the formulation of general descriptions of human behavior or in the explanation of abnormal mental conditions. (1970, 399; for similar reasons, see, e.g., P. S. Churchland 1986, 223) (Italics added)

From this it follows that our mentalist picture is insufficient for a general account of human behavior and cognition, though it is enough for daily transactions. However, we should desire a sufficient account. Then, the mentalist picture should be improved by unending revisions for the following reason:

The crude and incomplete causal theory embodied in commonsense psychology should not be expected to *survive* the next hundred years of central nervous system studies intact. It would be surprising if concepts like belief and desire found correspondents in a neurophysiological theory, considering how limited their explanatory and predictive power is, even for gross behavior. (1970, 399) (Italics added)

This passage reflects explicitly a powerful critique of folk psychology focusing on concepts of belief and desire. It emphasizes the explanatory limitations and predictive weaknesses of folk psychology regarding even gross behavior. It claims that future brain science would not match our current self-conception.

The physical behavior which, on Armstrong's analysis, a given intention is apt to cause, may be the product of causes whose complexity cannot be brought into even *rough correspondence* with the simple elements of a present-day psychological explanation. (1970, 399) (*Italics added*)

I can reasonably say that Nagel's objections are not directed against scientific materialism but against folk materialism. He says that the solution lies in "a more advanced theory of human functioning" as follows:

If that is so, then a physicalist theory of human functioning will not take the form of identifications between old-style psychological states and microscopically described physical states of the central nervous system. It will be couched instead in the concepts of *a more advanced theory of human higher functioning*. (1970, 399) (*Italics added*.)

Old psychological concepts will not work in the future. They will become archaic. In a future theory of cognition, we will need novel mental terms and a new objective phenomenological vocabulary.

I hope that this is a sufficient introduction to Nagel's ideas about physicalism, objectivity, and folk psychology. Now let us see which and in what ways philosophers misconceived his position about consciousness.

3. How is Nagel misconceived?

Thomas Nagel is probably one of the most-cited living analytic philosophers of the second half of the last century and arguably the most-cited philosopher of mind ever.⁵ His arguments are often cited as refuting some or all versions of physicalist theories (Lycan 2003, 186; Avramides 2006, 228–30; Wider 1990; Gorman 2006; Taylor 2016, 78; Thomas 2009, 35), as denying the possibility of giving a naturalistic/objective account of consciousness (Flanagan 1985, 373; Ratcliffe 2002, 353; Bergström 2009, 76; Stoljar 2017, sec. 16), or as showing that the arguments in favor of

⁵ For his 1974 paper, Web of Science (WoS) citations (as of June 17, 2020): 2,501 counts, from Clarivate.

physicalism are not cogent (Nagasawa 2003, 377). The only two works (in either one of the sections of the respective books) that have somewhat focused on the connection between Nagel's physicalism and his objective phenomenology are (Stubenberg 1998; Thomas 2009). Although Stubenberg argues that the objective phenomenology project is to clear the path for future physicalism (p. 42), Thomas argues that Nagel's nonphysicalism is compatible with his objectivism (p. 38).⁶ None of these accounts adequately addresses the relationship between physicalism, objectivity, and massive deficiencies of folk psychology, as Nagel construes them.

Nagel is largely a critical defender of objectivism (see Nagel 1986, 5; Nagel 1974, 449; for a defense of not objectivity but the scientific explanation of mental, see Nagel 2013). If objectivity is naturally associated with the third-person externalist viewpoint, subjectivity is associated with the first-person internal viewpoint. An objective point of view is "a progressive departure from earlier internal views" or subjective view (Boruah 1995, 339). They are not contrasting viewpoints, but "are part of a single spectrum of vision" (1995, 339). These two views are not mutually exclusive. This is why Nagel talks about "mental objectivity" (1986, chap. 2). Before discussing the relationship between his physicalism and the project of objective phenomenology, we should take a closer look at his conception of physicalism.

Nagel's earliest definition of physicalism is "I mean by physicalism the thesis that a person, with all his psychological attributes, is nothing over and above his body, with all its physical attribute" (1965, 339). He is "inclined to believe that some weak physicalist theory of the third type is true" and that "any plausible physicalism will include some state and event identities, both particular and general" (p. 340). The first type is identity theory, and the fourth is something even weaker than the token physicalism.⁷ His acknowledging the truth of physicalism is abductive. He has some reasons to believe that some sort of physicalism should be true. He gives no

⁶ Alan Thomas tends to interpret Nagel's objective phenomenology project in an expressly anti-physicalist manner. The problem is that he does not provide an argument for this and just assumes it.

⁷ We should not be perplexed by the labyrinthine complexities of the concept of the physical and of theories of physicalism. These are extraneous to my present discussion.

argument for this. In fact, his problem is not to defend or refute physicalism but just to defeat the then widespread arguments for the conclusion that physicalism must be false.

My attitude toward it is precisely the reverse of my attitude toward physicalism, which repels me although *I am persuaded of its truth*. The two are of course related, since what bothers me about physicalism is the thought that I cannot be a mere physical object, cannot in fact be anything in the world at all, and that my sensations and so forth cannot be simply the attributes of some substance. (1965, 356, cf. 1971, 111) (All but the last italics have been added)

Interestingly, from this passage, we see that Nagel has been a physicalist in as early as 1965; he was persuaded of its truth.⁸ Crane claims that the point for Nagel is that we cannot fully understand physicalism. However, this interpretation of Crane is problematic because Nagel does not say that "we cannot understand it ever." He does not claim that a physicalist account of consciousness cannot be given, only claims that nobody has yet given a plausible account. Thus, there remains a conceptual barrier in front of us.

What Nagel says is that when assuming the available mentalistic conception of human beings, the identity of mind and brain appears impossible to be true. On the other hand, he explicitly acknowledges that some weak form of physicalism is true. The reasonable conclusion thus is that we should revise and expand upon our available set of mentalistic ideas. This is why claiming that Nagel argues for the strict irreducibility of the mental to the physical is in erroneous. Nagel has only argued for conditional conceptual irreducibility given our self-conception, not for categorical irreducibility. Concepts reform, and categorical irreducibility disappears. The unintelligibility of the physicalist account of experience then ends.

In his atypical form of physicalism, the classical distinction between physical and mental becomes obsolete. The subjective-objective relationship

⁸ In conversation, many people asked me how I happened to be sure that Nagel did not substantially change his attitude toward physicalism in the last half a century. The answer to that question lies in my exposition of his replacement of the physicalism question with the problem of objectivity.

replaces the mental–physical dichotomy (Nagel 1979, 202). However, the notion of objectivity is importantly revised in stages as follows: “The development goes in stages, each of which gives a more objective picture than the one before” (1980, 79). If we can see that the question of physicalism is the problem of objectivity in guise, we can accept that the physicalism problem is not an ontological but a methodological one. This is so since “[O]bjectivity is a method of understanding” (1980, 77). The categories of subjectivity and objectivity replaces the categories of mental and physical. I think that this is key to understanding why Nagel occasionally refers to his approach as neutral monism or dual-aspect theory.

4. The problem of physicalism lies in giving an objective account of the subjective

“... the physical is a substitute for objectivity in posing the mind-body problem.” (Nagel 1979, 202)

Pär Sundström notes that many people reject the reading of Nagel 1974 to the effect that consciousness cannot be explained in physicalist terms:

In conversation, I have often met with the claim that Nagel does not try to argue that experience cannot be accounted for in physicalist terms, but merely illustrates an intuition. I think there is something true about this. (2002, 92)

Nagel asserts that “The mind-body problem exists because we naturally want to include the mental life of conscious organisms in a comprehensive scientific understanding of the world” (1993, 1). He “offers a defense and a critique of objectivity” (1986, 5).⁹ For Nagel, the core problem lies in how to give an ever increasingly objective account of the subjective.¹⁰

⁹ His critique of objectivism is limited to some ambitious claims of some natural scientists who venture fall beyond the scientific spirit and make bold assertions bolstered by a metaphysical worldview (Nagel, 2012, ch. 1). For him, those who choose anti-reductionism over objectivity deserve neither.

¹⁰ By far, the most elaborate version of his objective phenomenology project is presented in his *The View from Nowhere* (1986, chap. II). This does not determine

For Nagel, if something is physical, "it has to be objective" (1974, 449, fn 15, for more on this issue see also 1979, 202). That is, if we are to explain the mental in physical terms, we have to characterize it as something objective. On the other hand, what we need might be mental objectivity (1986, 17). Nonetheless, Nagel anticipates that in the future when relationships between the mental and physical are fully understood, "the fundamental terms" of the theory that explains that relation will not fall squarely with our current categories of physical or mental. That is, for Nagel, the physical account of the mental will remain improbable without "giving much more thoughts" to the general problem of the subjective and objective (1974, 450). In fact, Nagel, in one of his less known works, states that the problem of physicalism is just a substitute for the question of objectivity (1979, 202; for a parallel claim, see Stoljar 2017) as follows:

The physical is an ideal representative for the objective in general; therefore much obscurity has been shed on the problem by faulty analogies between the mental-physical relation and relations between the physical and other objective aspects of reality.

Nagel explores the connection between the physical and the objective. Having a more objective/detached account of consciousness is his desire (1980, 91) because "objectivity is naturally linked with reality" (1979, 202). If the internality of our psychology (i.e., the subjectivity of consciousness) is real, then there must be an objective account of it. Several central philosophical problems in the philosophy of mind are in fact the disguised expressions of the objectivity problem as described below.

As determinism is a substitute for externality or objectivity in posing the problem of free will, so the physical is a substitute for objectivity in posing the mind-body problem. All the disputes over causal role, theoretical identification, and functional realization, while of interest in themselves, fail to give expression to the central issue that makes the mind-body problem so hard. (Nagel 1979, 202)

whether the objective account of the subjective aspect of experience can be absolute or complete and whether it is desirable to achieve it maximally (Thomas 2009, 33).

What makes the problem of consciousness intractable, thus, is not that there is a mystery about how the physical gives rise to the mental but our lack of a suitable notion of objectivity. Our current notion of objectivity is confined to pure physical objectivity. It pushes the phenomenal aspect of experience out to the purely subjective side of the debate. The phenomenological aspect of experience should be made amenable to objective exploration. Nagel proposes doing this through his objective phenomenology project. The target of this project is “to clear the path for a future physicalism” (Stubenberg 1998, 42; also see Matthews 2009, 71). This is indeed the case:

Apart from its own interest, a phenomenology that is in this sense objective may permit questions about the physical basis of experience to assume a more intelligible form. Aspects of subjective experience that admitted this kind of objective description might be better candidates for objective explanations of a more familiar sort. (Nagel 1974, 449–50)

This is Nagel’s guess. In the future, it is possible to develop an objective phenomenological vocabulary to answer the question: “what is it like to be a bat for a bat?” (see Atkins 2013). Nagel does not deny the possibility of giving an objective account of consciousness. In contrast, he strives for this.

5. Concluding remarks

Finally, I must state that most of the things that philosophers say about Nagel have no basis at all in what Nagel actually says about the possibility of giving an objective account of the subjective aspect of experience. Thomas Nagel is not against physicalism as such, but he is against some mistaken forms of it. Nagel acknowledges the truth of weaker forms of physicalism. He does not deny the power of scientific achievements or objective methodology in the examination of philosophical problems, even including the subjective aspect of experience. He is not a subjectivist. Quite the reverse, he claims that we should pursue an unending inquiry to find the objective nature of subjective phenomena. The project of objective phenomenology is proposed for this aim. Nagel has shown us a way to conceive the

consciousness problem in an objective manner. Consciousness *is* something intractable, yet it can be made tractable in a physicalist framework through an objective phenomenology project. It is fallacious to demand a direct answer to such a complex problem as consciousness without first analyzing the basis of the question itself. This is what Nagel did. He challenged the widespread assumption that the problem of consciousness is intractable by its very nature.

References

- Allen-Hermanson, Sean. 2015. "Strong Neurophilosophy and the Matter of Bat Consciousness: A Case Study." *Erkenntnis* 80 (1): 57–76.
<https://doi.org/10.1007/s10670-014-9612-2>
- Alter, Torin. 2002. "Nagel on Imagination and Physicalism." *Journal of Philosophical Research* 27: 143–58. https://doi.org/10.5840/jpr_2002_28
- Atkins, Richard Kenneth. 2013. "Toward an Objective Phenomenological Vocabulary: How Seeing a Scarlet Red Is like Hearing a Trumpet's Blare." *Phenomenology and the Cognitive Sciences* 12 (4): 837–58.
<https://doi.org/10.1007/s11097-012-9288-5>
- Avramides, Anita. 2006. "Thomas Nagel: The View from Nowhere." In *Central Works of Philosophy*, edited by John Shand, 227–46. Buckinghamshire: Acumen.
- Bergström, Lars. 2009. "Thomas Nagel—Recipient of the Rolf Schock Prize in Logic and Philosophy, 2008." *Theoria* 75 (2): 76–78.
<https://doi.org/10.1111/j.1755-2567.2009.01033.x>
- Bickle, John. 1992. "Revisionary Physicalism." *Biology and Philosophy* 7 (4): 411–30. <https://doi.org/10.1007/BF00130060>
- Bickle, John. 1998. *Psychoneural Reduction: The New Wave*. Cambridge, Mass.: MIT Press. A Bradford Book.
- Bond, E. J. 2005. "Does the Subject of Experience Exist in the World?" *Philosophy and Phenomenological Research* 71 (1): 124–124.
<https://doi.org/10.1111/j.1933-1592.2005.tb00433.x>
- Boruah, Bijoy. 1995. "The Non-Rational Foundation of Moral Reality." In *The Philosophy of P.F. Strawson*, edited by Pranab Kumar Sen and Roop Rekha Verma, 327–45. New Delhi: Indian Council of Philosophical Research.
- Churchland, Patricia Smith. 1986. *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge, MA: MIT Press.
- Crane, Tim. 2007. "Cosmic Hermeneutics vs. Emergence: The Challenge of the Explanatory Gap." In *Explaining the Mental: Naturalist and Non-Naturalist*

- Approaches to Mental Acts and Processes*, edited by Carlo Penco and Michael Beaney, 22–34. Newcastle: Cambridge Scholars Publishing.
- D’Oro, Giuseppina. 2007. “The Gap Is Semantic, Not Epistemological.” *Ratio* 20 (2): 168–78. <https://doi.org/10.1111/j.1467-9329.2007.00355.x>
- Dennett, Daniel C. 1991. *Consciousness Explained*. New York: Back Bay Books.
- Flanagan, Owen. 1985. “Consciousness, Naturalism, and Nagel.” *Journal of Mind and Behavior* 6 (3): 373–90. <https://www.jstor.org/stable/43853176>
- Foss, Jeffrey. 1993. “Subjectivity, Objectivity and Nagel on Consciousness.” *Dialogue* 32 (04): 725–36. <https://doi.org/10.1017/S0012217300011367>
- Gorman, Michael. 2006. “Nagasawa vs. Nagel: Omnipotence, Pseudo-Tasks, and a Recent Discussion of Nagel’s Doubts About Physicalism.” *Inquiry* 48 (5): 436–47. <https://doi.org/10.1080/00201740500241896>
- Lycan, William G. 2003. “Philosophy of Mind.” In *The Blackwell Companion to Philosophy*, edited by N. Bunnin and E. P. Tsui-James, 2nd ed., 173–201. London: Blackwell.
- Matthews, Lucas J. 2009. “What Is It like to Be a Dualist? On the Real Challenge of Nagel’s 1974 Paper.” *Indiana Undergraduate Journal of Cognitive Science* 4: 71–80.
- McGinn, Colin. 1980. “Philosophical Materialism.” *Synthese* 44 (2): 173–206.
- McGinn, Colin. 1989. “Can We Solve the Mind–Body Problem?” *Mind* 98 (391): 349–66. <https://doi.org/10.1093/mind/XCVIII.391.349>
- McHenry, Leemon B., and George W. Shields. 2016. “Analytical Critiques of Whitehead’s Metaphysics.” *Journal of the American Philosophical Association* 2 (3): 483–503. <https://doi.org/10.1017/apa.2016.21>
- Nagasawa, Yujin. 2003. “Thomas vs. Thomas: A New Approach to Nagel’s Bat Argument.” *Inquiry* 46 (3): 377–94. <https://doi.org/10.1080/00201740310002415>
- Nagel, Thomas. 1965. “Physicalism.” *Philosophical Review* 74 (3): 339–56. <https://doi.org/10.2307/2183358>
- Nagel, Thomas. 1970. “Armstrong on the Mind.” *Philosophical Review* 79 (3): 394–403. <https://doi.org/10.2307/2183935>
- Nagel, Thomas. 1971. “Physicalism,” with ‘Postscript.’ In *Materialism and the Mind–Body Problem*, edited by D. Rosenthal. Englewood Cliffs, N.J.: Prentice-Hall.
- Nagel, Thomas. 1974. “What Is It like to Be a Bat?” *Philosophical Review* 83 (4): 435–50. <https://doi.org/10.2307/2183914>
- Nagel, Thomas. 1979. “Subjective and Objective.” In *Mortal Questions*, Cambridge, Mass.: Cambridge University Press. 196–213.
- Nagel, Thomas. 1980. “The Limits of Objectivity.” In *Tanner Lectures on Human Values*, 77–139. Accessed November 24, 2020. https://tannerlectures.utah.edu/_documents/a-to-z/n/nagel80.pdf

- Nagel, Thomas. 1986. *The View From NoWhere*. Oxford and New York: Oxford University Press.
- Nagel, Thomas. 1993. "What Is the Mind-Body Problem?" In *Experimental and Theoretical Studies of Consciousness*, edited by G.R. Block and J. Marsh, 1–7; discussion 7–13. Chichester: John Wiley and Sons.
- Nagel, Thomas. 1997. *The Last Word*. New York and London: Oxford University Press.
- Nagel, Thomas. 1998. "Conceiving the Impossible and the Mind-Body Problem." *Philosophy* 73 (285): 337–52. <https://doi.org/10.1017/S0031819198000035>
- Nagel, Thomas. 2002. "The Psychophysical Nexus." In *Concealment and Exposure: And Other Essays*, 194–235. New York, N.Y.: Oxford University Press.
- Nagel, Thomas. 2009. "The I in Me: I and Me." *London Review of Books*, 31 (21). <https://www.lrb.co.uk/the-paper/v31/n21/thomas-nagel/the-i-in-me>
- Nagel, Thomas. 2012. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*. Oxford and New York: Oxford University Press.
- Nagel, Thomas. 2013, August 18. "The Core of 'Mind and Cosmos.'" *The New York Times*. <https://nyti.ms/2k0ZOyT>
- Nagel, Thomas, Anne Erreich, Richard J. Kessler, Barry Rand, and Jerome Wakefield. 2016. "An Exchange with Thomas Nagel: The Mind-Body Problem and Psychoanalysis." *Journal of the American Psychoanalytic Association* 64 (2): 389–403. <https://doi.org/10.1177/0003065116647053>
- Pernu, Tuomas K. 2017. "The Five Marks of the Mental." *Frontiers in Psychology* 8: 1–19. <https://doi.org/10.3389/fpsyg.2017.01084>
- Ratcliffe, M. 2002. "Husserl and Nagel on Subjectivity and the Limits of Physical Objectivity." *Continental Philosophy Review* 35 (4): 353–77. <https://doi.org/10.1023/A:1023934903646>
- Seager, William. 2014. "Why Physicalism?" *Mind and Matter* 12 (2): 1–28.
- Stoljar, Daniel. 2017. "Physicalism." *The Stanford Encyclopedia of Philosophy*. (Winter 2017 Edition), edited by Edward N. Zalta. Last updated October 18, 2017 <https://plato.stanford.edu/archives/win2017/entries/physicalism/>
- Stubenber, Leopold. 1998. *Consciousness and Qualia*. Amsterdam: John Benjamins Publishing.
- Sundström, Pär. 2002. "Nagel's Case against Physicalism." *SATS* 3 (2): 91–108. <https://doi.org/10.1515/SATS.2002.91>
- Sundström, Pär. 2018. "How Physicalists Can—and Cannot— Explain the Seeming 'Absurdity' of Physicalism." *Philosophy and Phenomenological Research* 97 (3): 681–703. <https://doi.org/10.1111/phpr.12394>
- Tartaglia, J. 2016. "What Is at Stake in Illusionism?" *Journal of Consciousness Studies* 23 (11–12): 236–55.

- Taylor, Elanor. 2016. "Explanation and the Explanatory Gap." *Acta Analytica* 31 (1): 77–88. <https://doi.org/10.1007/s12136-015-0260-1>
- Thomas, Alan. 2009. *Thomas Nagel*. Buckinghamshire: Acumen.
- Wider, Kathleen. 1990. "Overtones of Solipsism in Thomas Nagel's 'What Is It Like to Be a Bat?' And the View from Nowhere." *Philosophy and Phenomenological Research* 50 (3): 481–99. <https://doi.org/10.2307/2108160>

Epistemic Foundations of Salience-Based Coordination

Vojtěch Zachník*

Received: 4 April 2019 / Revised: 19 October 2020 / Accepted: 11 November 2020

Abstract: This paper aims to assess current theoretical findings on the origin of coordination by salience and suggests a way to clarify the existing framework. The main concern is to reveal how different coordination mechanisms rely on specific epistemic aspects of reasoning. The paper highlights the fact that basic epistemic assumptions of theories diverge in a way that makes them essentially distinctive. Consequently, recommendations and predictions of the traditional views of coordination by salience are, in principle, based on the processes related to the agent's presumptions regarding the cognitive abilities of a co-player. This finding implies that we should consider these theories as complementary, and not competitive, explanations of the same phenomenon.


Keywords: Coordination; correct belief; epistemic symmetry; rationality; salience.

1. Introduction

There are many coordination challenges in our everyday lives (greeting patterns, traffic rules, dancing moves, etc.), yet we do not feel that these

* University of Hradec Králové

 <https://orcid.org/0000-0002-3311-8756>

 Faculty of Philosophy, University of Hradec Králové, Rokitanského 62, 500 03
Hradec Králové, Czech Republic

 vojtech.zachnik@uhk.cz



kinds of everyday interactions involve some sort of obstacle because our behaviour usually seems straightforward and effortless. The key question is how this behaviour emerges for the first time if we cannot rely on precedent, agreement, or rules. What are the underlying processes enabling this type of interdependent behaviour of multiple agents? How many different but parallel ways can bring about coordination in this setting? The notion of *saliency* was preliminarily specified in terms of “standing out” or “conspicuousness” (Schelling 1960; Lewis 1969a),¹ and it was used to explain the process of inducing coordinated actions of agents who are not able to appeal to any stronger background or decision principle.² Namely, an individual who wants to coordinate with others, but does not know which behavioural pattern is precisely suitable for the situation, may look for the assistance of a salient feature of an interaction (contextual clue, labelling of choice) and then coordinate (Schelling 1960).

Currently, the issue has been revived as the topic of the emergence of coordination is reflected by the new empirical evidence (Mehta et al. 1994; Bardsley et al. 2010; Colman et al. 2014). In general, my aim is to assess two leading proposals and answer the following question: Is it the case that people coordinate by saliency because they frame contextual cues and conceive the situation from a new perspective—as described by the *variable frame theory* (Bacharach and Bernasconi 1997; Bacharach and Stahl 2000; Bacharach 2006), or because individuals have a ‘hunch’ about another’s behaviour and try to respond to this as best as possible—as suggested by the *cognitive hierarchy theory* (Camerer and Chong 2004)? I want to argue that both theories are built upon some elementary assumptions about the beliefs of others, and the logical structure of these epistemic foundations makes these two approaches compatible. Therefore, I suggest there are two parallel kinds of saliency-based coordination processes (based on the above-

¹ Alternatively, Sugden (2011) defines saliency more clearly as “an individual’s pre-reflective perception that certain elements of the situation stand out from the rest”.

² There are three widely accepted ways of how coordination can emerge (Lewis 1969, 24–42). The first two are implicit and not purposeful: coordination based on saliency, the one I am discussing in the paper, and coordination due to precedent (Schelling 1960, chap. 4; Young 1996). The third is an explicit communication such as agreement.

mentioned theories), and their usage is determined by the epistemic context of an interaction. More specifically, I will show that *epistemically symmetrical* conditions of interaction favour reasoning modelled by variable frame theory, whereas *epistemically asymmetrical* conditions support reasoning with cognitive hierarchy.

The paper is structured as follows: In Section 2 I reveal the problem of coordination, and why it presents a challenge for any theory of coordination. Section 3 discusses two recent and dominant views of saliency which are then described and analyzed in terms of their epistemic underpinnings in Section 4. Finally, Section 5 summarizes my argument by introducing notions of epistemic symmetry and asymmetry and reflects some general implications.

2. Coordination Problem

In the opening paragraph, I have briefly mentioned cases in which people try to coordinate their behaviour to achieve a mutually beneficial outcome. For example, a pedestrian and a car driver face the crucial decision of whether to *wait* or *go* once they reach an intersection at the same time (assuming there are no traffic lights). Undoubtedly, their goal is to arrive at a state of the world in which the only one chooses to *go*, while the other *waits*. Similarly, when we meet someone in a theatre, we tend to greet this person. But how? Should we hug a person, kiss her, shake her hand? In game theory, it is common to use a notion of coordination game to denote a strategic interaction that poses an issue of selection between many viable alternatives. The game represents a situation in which two or more agents make a decision from the set of available options with the intention of directing their actions towards a certain outcome. Moreover, agents' preferences in the interaction are such that they favour the same outcome (since it is beneficial for both of them), and that outcome cannot be achieved by acting alone or by neglecting to consider others' actions. Because of this, each player forms beliefs about other agents' actions to estimate the potential consequences.³ It is also important to keep in mind that the preference

³ In other words, coordination as a type of strategic interaction expresses the idea that beliefs about others have an important significance within this decision-making

coincidence is a central feature for distinguishing coordination from situations with some degree of conflict, such as zero-sum games or mixed-motive games.⁴

In order to generalize these ordinary intuitions and to abstract structural properties from any contingent features, I build a simple game-theoretical model of a prototypical coordination interaction. For the sake of simplicity, assume a two-player pure coordination matching game (more specifically, a one-shot non-cooperative normal form game with perfect information):

$$\Gamma = (N, S, U)$$

Set of players: $N = (i, j)$

Set of Strategies: $S = (s_1, s_2, s_3) \dots$ for i and $j \in N$.

Payoffs: Let $u_i(s_i, s_j)$ denote player i 's payoff given her strategy s_i and co-player's strategy s_j .

U: $U(s_i, s_j) \dots$ if $s_i = s_j$ then $u_i = u_j = 1$
 \dots if $s_i \neq s_j$ then $u_i = u_j = 0$

Based on this formal game-structure, it is easy to represent the interaction by a payoff-matrix, because the formal n -tuple, Γ , contains all the information necessary for such a move.

	s_{j_1}	s_{j_2}	s_{j_3}
s_{i_1}	1, 1	0, 0	0, 0
s_{i_2}	0, 0	1, 1	0, 0
s_{i_3}	0, 0	0, 0	1, 1

Figure 1. Pure Coordination game

process. This aspect makes it different from decisions such as whether I should take an umbrella today if I suspect that it might rain, or, when buying a new car, the consideration of such factors as fuel consumption and safety.

⁴ Schelling (1960) suggested a concept of a continuum of interactions with two extremes on either side: pure coordination (agents' preferences perfectly coincide), and pure conflict (preferences are directly opposed).

You can see that the players decide simultaneously what to do, and then they establish an outcome of the game, a product of their strategy-combination (chosen from row and column by respective players). Given this, we can easily and clearly evaluate their choices, and say how preferable the result is considering their utilities.⁵

However, what complicates this situation is a feature of the game that allows three possible outcomes to be equally acceptable (ranked by both player with utility 1).⁶ This is *the coordination problem*. It is a problem because it questions our competence to decide what exactly the solution of the game is, and therefore it imposes the difficulty of selection (Harsanyi and Selten 1988). In real-life situations, people usually use *agreement* to solve the indeterminacy, or they rely on some *precedent* that helps to stabilize their expectations of a possible solution. On the other hand, under circumstances when agents cannot communicate and no pattern of behaviour from previous encounters is known, there is still one remaining way to solve this curse of symmetry. External factors can make one of the choices somehow salient. Intuitively speaking, salience breaks the symmetry by the fact that some strategies will stand out and appear strikingly different in comparison to others.

3. Theories of Salience

The focus of interest for the rest of the paper, therefore, is a particular model of interactions that are well represented in everyday life, and the role

⁵ The assumptions behind this are the standard ones: preferences are expressed in Von Neumann-Morgenstern utilities, and players are rational in the sense that they maximize expected utility due to common knowledge assumptions.

⁶ In game-theoretical terms, the game has three payoff-symmetric pure Nash equilibria. In addition to these, there is also Nash equilibrium in mixed strategies. For more details on categories of coordination games, see Camerer (2003). Also, it is worth mentioning that the many-solutions problem is crucial for the game because the coincidence of interests is not a sufficient condition for coordination, as shown by Ullmann-Margalit (1977, 79–80).

of salience in these interactions.⁷ However, the phenomenon of salience (as preliminarily mentioned in Section 1) is, for now, placed in a more refined theoretical foundation that provides more robust explanatory grounds for reflection on different modes of reasoning. I will briefly introduce two prominent theories of salience-based coordination—*variable frame theory*, and *cognitive hierarchy theory*—and then reveal their epistemic background to show why I think these theories complement each other.

Michael Bacharach invented a formal extension of the game-theoretical model of coordination in *variable frame theory* (Bacharach and Bernasconi 1997; Bacharach and Stahl 2000; Bacharach 2006).⁸ It attempts to explain a salient choice in terms of conceptual frames or labelling, given that, it takes a step beyond orthodox game theory by enriching apparatus with the notion of frame.⁹ Frame is a set of concepts or labels by which the agent perceives the interactive situation, and Bacharach accounts for a salient choice in terms of the agent's framed decision-making. If individuals conceptualize interaction through the same lens, then salience may occur and guide their decisions towards coordination. Suppose you are playing a matching game with another person and you must choose from a set of five distinct objects: whisky, wine, water, beer, or sherry (assume they have stickers attached to them, so you do not have to taste them). Variable frame theory assumes that players describe the interaction through various predicates. For instance, *alcoholic* and *non-alcoholic* suggest themselves as

⁷ Technically speaking, all attention is devoted to a one-shot normal form coordination game in which the preferences of agents perfectly coincide and no communication or (direct) past experience is allowed. Unless stated otherwise, in all remaining sections I consider such a game as the default option.

⁸ Bacharach followed in the footsteps of Gauthier (1975), and significantly extended his original intuition by providing a comprehensive theoretical framework. It was Gauthier who first innovatively suggested that salience induces a payoff-modification that transforms the original pure coordination game into a game with asymmetric equilibria (Hi-Lo game); and he augmented the account by the principle of coordination, which states the normative claim for an agent to choose a payoff-dominant equilibrium.

⁹ In traditional game theory, it does not matter how an agent perceives a game since theorists have an objective way to describe interaction by indicating strategies and payoffs. See Luce and Raiffa (1957).

obvious categories. The non-alcoholic drink (water) appears to be a good candidate for the salient choice (cf. Bardsley et al. 2010).

Why? To detail how this process of framing works, I show the formalization of the example in conformity with Bacharach (2006). The standard model of game is extended with frame F containing different families of predicates (here, generic family F_0 and F_1) and these families are, moreover, specified by parameters of how likely their occurrence is in the player’s mind (availability, $v(F) = p$; $0 \leq p \leq 1$), and what strategies—in the sense of traditional objective game—are among the predicates in the family (extension of a predicate, E). As a result of this configuration, it is possible to capture the idea of how salience transforms the payoff structure of the game.

$$\begin{aligned} \Gamma_{\text{beverages}} &= (N, S, U, F) \\ N &= (i, j), S = (\text{whisky, wine, water, beer, sherry}) \\ U: U &(s_i, s_j) \dots \text{if } s_i = s_j \text{ then } u_i = u_j = 1 \\ &\dots \text{if } s_i \neq s_j \text{ then } u_i = u_j = 0 \\ F_1 &= \{F_0, F_1\}, E: F \rightarrow S \\ F_0 &= \{\text{thing}\}; E(\text{thing}) = \{\text{whisky, wine, water, beer, sherry}\}, \\ &v(F_0) = 1 \\ F_1 &= \{\text{alcoholic, non-alcoholic}\}; v(F_1) = 1 \\ E(\text{alcoholic}) &= \{\text{whisky, wine, beer, sherry}\}, \\ E(\text{non-alcoholic}) &= \{\text{water}\} \end{aligned}$$

	<i>Choose the non-alcoholic</i>	<i>Pick an alcoholic</i>	<i>Pick a thing</i>
<i>Choose the non-alcoholic</i>	1, 1	0, 0	0.2, 0.2
<i>Pick an alcoholic</i>	0, 0	0.25, 0.25	0.2, 0.2
<i>Pick a thing</i>	0.2, 0.2	0.2, 0.2	0.2, 0.2

Figure 2. Framed matching game

In comparison to all other possibilities, the *non-alcoholic* drink offers the best chance for a match in coordination, as the extension of the predicate contains only one element. This conclusion is not very surprising, but the

formal notation reveals further details.¹⁰ First, an important consequence of framed games is that they break the curse of symmetry by introducing asymmetrical equilibria (in bold in Figure 2).¹¹ Therefore, the coordination problem is essentially altered into a new game structure (Hi-Lo game) solely under the influence of framing. Second, this reshaping of the original pure coordination game brings us closer to an explanation. However, it does not—in this form—establish a definite solution to the game. The reason for this is simple: a player may be equally justified in ‘picking an alcoholic drink’ if he or she reasonably expects the other to make the same decision.

The missing piece of the puzzle is the kind of reasoning that supports this decision. Could a player appeal to some principle of coordination when choosing a Hi-equilibrium (in my example ‘choosing the non-alcoholic drink’)? Variable frame theory answers this question by explaining this mechanism and providing its elaborate justification. Rational agents do not consider their choices only on the basis of standards of individual rationality—they think as a team, considering what is beneficial for *them*, collectively.

This means that a new principle of rationality enters the scene, with the formal consequence of directing individuals’ choices towards a Pareto-optimal equilibrium.¹² Bacharach is convinced that there are strong reasons that lead individuals to team-beneficial choices since we tend to identify with a certain group. In particular, group identification occurs as a result of

¹⁰ Keep in mind that both theories I present provide a formal explanation of salience, therefore they make no further assumptions as to what specific factors trigger this effect. Moreover, I believe that every substantive theory would be incomplete in its content since it is difficult to list all significant building blocks. And of course, the vast diversity of cultural contexts makes this effort even harder.

¹¹ What happens in special cases when there are many singleton predicates or availability of families of predicates variables is not an important issue here because it does not weaken my conclusion. For a complete account, see Bacharach and Bernasconi (1997) and Bacharach (2006).

¹² Strategy combination (s_i, s_j) is Pareto-optimal if there is no other combination $(s_i^*, s_j^*) \in S$ that satisfies:

- a) $\forall i \in N, U(s_i^*, s_j^*) \geq U(s_i, s_j)$
- b) $\exists i \in N, U(s_i^*, s_j^*) > U(s_i, s_j)$

perceived common interest, or strong interdependence (Bacharach 2006, 142–44), and is followed by the mode of team reasoning (or we-reasoning).¹³ Variable frame theory, broadly speaking, establishes a new form of non-standard reasoning that revises the traditional conception of rationality such that agents are now considered to be capable of recognizing an efficient outcome that gives *them* the best prospect of coordinating.¹⁴ Therefore, salient choice is produced by a particular frame that transforms a game-structure from pure coordination to Hi-Lo game, and then agents make a decision as team members in favour of a mutually beneficial outcome. These two essential components of the theory—framing and team reasoning—reliably explain why ‘water’ is the salient choice in the prototypical matching game.

But, to make things more complex, there is another alternative explanation for coordination by salience that stems from *cognitive hierarchy theory* (Camerer and Chong 2004).¹⁵ This model is, in principle, based on the assumption of a boundedly rational agent, who takes a limited number of reasoning steps before he or she decides, and, in the case of coordination, whose strategic thinking is rooted in some kind of rudimentary non-strategic non-rational salience. Hence, an overall account of salience-based coordination presented by this theory rests on two pillars: one that establishes weak symmetry-breaking behaviour, and the second, which postulates a finite belief hierarchy, and an individual who chooses the best strategic response

¹³ Some have identified several problems with team reasoning. It seems to be too narrowly specified in terms of social categorization (Hindriks 2012), and somewhat unstable in experimental testing either due to slight payoff asymmetry (Crawford, Gneezy, and Rottenstreich 2008) or due to the influence of other strategic options (Cooper et al. 1990).

¹⁴ This revisionary standpoint is, however, highly controversial as it puts into question a standard assumption of game theory—methodological individualism. Yet, on the other hand, many experimental studies show evidence in favour of this reasoning mode (Bacharach and Bernasconi 1997; Colman, Pulford, and Rose 2008; Bardsley et al. 2010).

¹⁵ The theory was initially developed to provide another way of thinking about solution concepts in game theory. And it had an impact on dominance-solvable games (e.g., Beauty Contest Game), but it also provides an interesting framework for thinking about coordination problem in a new way.

with regard to his or her beliefs.¹⁶ But how do these two features fit together? The whole result of salience proceeds in two steps that are captured and formalized by the theory. Initially, there is a non-uniform probability distribution on the set of strategies, sometimes known as *primary salience* (Mehta, Starmer, and Sugden 1994), whose role is to disrupt the symmetry of the coordination problem. A natural interpretation of this could be that agents have a psychological propensity to pick some strategies without any incentive and in the absence of strategic reasoning. Such behaviour then exhibits randomization over some strategies that might be particular to a certain cultural background, contextual information, or simply because of their uniqueness and conspicuousness. This part of the story might be sufficient to explain “picking behaviour”, or why there is a concentration of choices in aggregate, as supposed by Schelling (1960), and tested by Bardsley et al. (2010). However, the presence of coordination success on many occasions demands a fuller explanation.

At this point, the second element, belief hierarchy, becomes involved. The theory introduces agents of various levels of reasoning who have the cognitive ability to recognize lower-level agents, and to choose the best strategy (best-response)¹⁷ given their assumptions about other players and their choices (Stahl and Wilson 1995; or Haruvy and Stahl 2007). Therefore, it attempts to establish an apparatus whose expressivity allows us to grasp the intuition that agents have a certain depth of reasoning and the cognitive ability to understand other minds (Ohtsubo and Rapoport 2006). In this respect, there are categories of agents depending on how many steps have been taken, or let us say that each of them is assigned a certain level of reasoning. For instance, a level 0 player (or L0) lies at the bottom of the hierarchy and has no mental model of other players’ choices or reasoning abilities. His behaviour in coordination is fully characterized by the non-

¹⁶ The origins of this idea can be traced back to Lewis (1969b, 24–36) and his account that coordination is feasible only by means of a system of high-order expectations.

¹⁷ The best-response decision rule is essentially based on the strict dominance principle—a core element of decision-making. It assumes that a rational agent can eliminate all strategies that are, in all respects, worse than the other options.

uniform probability distribution p_0 (i.e., primary salience).¹⁸ Next, level 1 represents an agent who is convinced that all her co-players are L0, and she decides to maximize her chances of a match with the others. The intuitive understanding of the model is that the level 1 player tries to guess what the most probable choice is according to his peers, influenced by primary salience, and then chooses the best response to that behaviour. The same holds for all higher-level reasoners: they expect others to be lower-level agents, and they have particular beliefs about the frequency of these types in the population. In other words, these beliefs express the probability of an encounter with a given type. For example, a level 2 player believes that he may coordinate with someone who is either L1 or L0, and acts in order to maximize his expected utility in anticipation of the respective behaviour of his co-player.

The specific implications of cognitive hierarchy theory for coordination games is straightforward: players—depending on their type—maximize their chance of compliance with others based on generally recognized primary salience (or picking behaviour), which successively leads to a concentration of choices around one of the equilibria. This dynamic process of reasoning sooner or later selects one of the possible alternatives with the support of original non-uniform probability distribution. As an example, consider the familiar game with beverages except now I will analyse it is using the apparatus of cognitive hierarchy theory. The first obstacle emerges with the issue of how to determine the likelihood of choosing a drink, i.e. p_0 . The traditional answer is that we do not have to specify this *a priori* because, essentially, it is a matter of empirical research. The aim of the theory of salience-based coordination is not to enumerate all the sources of salience, but, rather, to demonstrate the formal consequences leading to one solution. Thus, as someone who lives in a country with a famous beer-drinking culture, I will imagine that primary salience in this case

¹⁸ One idealized assumption is that p_0 distribution is for all agents (even for higher levels) the same. The reason for such simplification is as follows: if theorists want to model salience-based behaviour then they think that the contextual background is commonly shared. Even though this might seem restrictive, as some individuals might display minor variations, the underlying idea is correct, at least for the instrumental purposes of the theory.

highlights one of the alcoholic drinks. Whatever an individual's choice is, let me assume a stable, reliable, and population-wide pattern of primary salience, for instance $p_0 = \{0.15 \text{ whisky}, 0.2 \text{ wine}, 0.1 \text{ water}, 0.4 \text{ beer}, 0.15 \text{ sherry}\}$. This describes a feature of what some individual picks if he or she would not consider others, but simply follow non-reflective inclinations. Although this behaviour occurs in coordination, the theory predicts some agents will strategize and focus on limited strategic thinking. A level 1 agent expects her co-player to behave in accordance with primary salience, and therefore she will choose pure strategy $s_{11} = \{\text{beer}\}$ since it gives her the highest expected utility ($u_{11}(s_{11}, p_0) = 0.4$). A Level 2 player believes that he may encounter either level 1 or level 0 with corresponding probabilities q_1 and q_0 (where $q_1 + q_0 = 1$), and he also forms beliefs concerning their behaviour (s_{11}, p_0). But how should and will a boundedly rational L2 agent act? Even if he imagines the scenario in which his co-player is certainly either type 1 or type 0 (i.e., $q_0 = 1$ or $q_1 = 1$), his best strategy is always to choose $s_{12} = \{\text{beer}\}$.¹⁹ Therefore, cognitive hierarchy theory describes coordination behaviour in this interaction as a gradual increase of the concentration of choices around one specific alternative.

To summarize, cognitive hierarchy theory explains salience by other means. Coordinating behaviour emerges rather as the result of the expectation of which option is most likely to be selected (given the various types of agents who may or may not think strategically). It is accepted that some players might be choosing blindly, but, overall, coordination is a result of a convergence of choices (in the example, it is convergence to the most popular drink). In comparison to variable frame theory, agents do not have to think as team members, and salience does not create a direct structural transformation. But let us pause for a moment and think more about what the analysis of the beverage game further reveals. The case clearly demonstrates the somewhat disturbing and striking result of divergent predictions provided by each of these theories in the very same game-setting. Whereas one theory ends with the selection of water, the second would suggest beer, and the question—*What would you like me to drink?*—seems to have no definite answer for now.

¹⁹ Expected utility for L2 player is $u_{12}(\text{beer}) = q_1 + 0.4q_2$, therefore $1 \geq u(s_{12}) \geq 0.4$.

Of course, in many other cases, the practical implication of the models would be identical, since one needs no more than to suppose that primary salience points in the same direction as a particular frame, keeping in mind that theoretical explanations and underlying assumptions differ (Bacharach and Stahl 2000). But the value of test cases such as the game with beverages rests more on the promise of assessing experimentally which theory is supported by the data, and the identification of a correct explanatory model. Unfortunately, the alleged behavioural litmus test did not provide results as promising as had been expected (Mehta, Starmer, and Sugden 1994; Bardsley et al. 2010; Colman, Pulford, and Lawrence 2014), and the problem seems to be, rather, that two parallel modes of reasoning are possible, and may influence individuals' decision-making in this type of situation. The question, then, is how to reconcile these dual processes?

4. Correct beliefs and belief in rationality

My proposal for a solution will be based on the idea that modes of reasoning in coordination with salience sustain certain epistemic standards which must be implicitly recognized by the interacting agents.²⁰ Hence, even before it comes to establishing the coordination outcome, every involved and strategically thinking agent makes some estimates concerning possible interaction scenarios, his or her co-player's behaviour, and beliefs (similarly Janssen 2001). Therefore, one can consistently claim that variable frame and cognitive hierarchy theory together provide an explanation of the coordination problem because each theory relies on different epistemic standards. In a nutshell, different epistemic background induces a distinctive coordination mechanism.

As we have seen with cognitive hierarchy theory, this approach of restricted reasoning belongs to a broader class of theories known as bounded rationality. And, as such, it makes rather less demanding epistemic assumptions, which are embedded into the concept of agent. First of all, cognitive

²⁰ In this spirit, I follow in the footsteps of the established programme of epistemic game theory, aiming to clarify solution concepts and their underlying epistemic principles. See more on this in de Bruin (2009) or Perea (2012).

hierarchy theory violates *correct beliefs assumption*. This is an important point in my argument I will specify later, but for now, it is sufficient to say that violation of correct beliefs means that even if coordination has been achieved by salience, we cannot say that individuals have correct beliefs about themselves. Thus, an inevitable consequence of cognitive hierarchy model is that every strategically thinking agent is, in fact, acting rationally (he decides for the best option with regard to his or her beliefs and level), even though he has an incorrect belief about his partner in coordination (there is an epistemic disharmony between individuals). The problem, therefore, is deeply rooted in the fact that an agent believes that a co-player is systematically mistaken in what *behaviour* the co-player attributes to *the agent himself*.

To illustrate this point, let me use the before-mentioned beverage-choosing game (Section 3). For instance, imagine a situation of two friends, John and Isaac, who want to order the same drink in a crowded bar with loud music. The only thing that matters is to have the same drink because they do not want to drink more than one type of beverage. Unfortunately, they have been separated by the crowd and each has to make an order independently of the other's decision. They face a typical coordination problem. How can they solve it? Cognitive hierarchy theory predicts that each will choose or pick a drink depending on his cognitive level. The crucial aspect now, however, is what beliefs they have about each other. Let say that John believes that Isaac will choose beer because he believes that John himself is randomly picking one of the drinks, and beer seems like the most attractive option (primarily salient). Given that, John chooses beer too, though he does that knowing that Isaac is mistaken about his actual beliefs. Remember, John believes that Isaac thinks that he is randomly picking. Coordination in this case will be successful despite the obvious epistemic discord.

Now, I will illustrate the issue of incorrect beliefs more formally, which allows me to capture this feature of the theory in a precise manner. I will assume that both agents (John and Isaac) are of the same level, say L2 players. Both expect that the partner will be L1 or L0 (with respective probabilities); and if their partner is L1, L2 player will also think that the co-player (as a L1 agent) has some beliefs about him, namely that L1 will think she is paired with an L0 player. However, we need to know not just

the agent's relevant level, but also information about his strategies. Then, it may be useful to represent formally by means of a simple notation such as t_{i2}^{beer} that each individual has a certain epistemic type for a given depth of reasoning and chosen strategy; t_{i2}^{beer} indicates that Isaac's (agent i 's) epistemic type is specific for an L2 player who is choosing the strategy *beer*.²¹ Epistemic type in a nutshell is a convenient way to express the beliefs an individual has, and how they are structured. Bearing this in mind, it is not very difficult to describe an agent's type for the game as follows:

$$t_{i2}^{\text{beer}}: t_{j1}^{\text{beer}} \rightarrow t_{i0}^{\text{alcoholic}}$$

$$t_{j2}^{\text{beer}}: t_{i1}^{\text{beer}} \rightarrow t_{j0}^{\text{alcoholic}}$$

In this display, you can see a case of how two agents of the very same level, on the one hand, have false expectations regarding the other player's level. Isaac (t_{i2}^{beer}) believes that John is L1, whereas, in fact, he is t_{j2} . This trivial result, though, can be easily avoided simply by stipulating that John is actually t_{j1} , and then it would prevent this type of incorrectness, which is not my direct concern here. On the other hand, a much more important implication of the model lies in what Isaac (t_{i2}) thinks about John's expectations about him. As previously stated, t_{i2} believes that he is interacting with John of L1.²² Or, more precisely, he believes that co-player j is a t_{j1} player who chooses *beer* because John expects that Isaac is an L0, who randomizes amongst alcoholic drinks (in accordance with p_0) and has no model of his co-players. And conversely, the same holds for t_{j2}^{beer} . Thus, a crucial consequence of cognitive hierarchy theory is that the agent (t_{i2}) assumes that his co-player is fundamentally wrong in her belief about how he will behave. This kind of incorrectness is different from the first, concerned with the hypothetical versus the actual level of the agent, and, furthermore, is a profound feature of the theory that helps outline its epistemic coordination roots.

²¹ Type-space notation is suitable here for the reason that it allows us to comprehend an aspect of nested beliefs in a simple and elegant way. Cf. Sillari (2008) or Geanakoplos (1992).

²² This is a harmless simplification, as I assume that $q_1=1$.

An important lesson here is that such description reveals a general tendency built into the theory: a rational agent (of sufficiently high level)²³ believes that the partner is fundamentally mistaken in his expectation about whom he is interacting with. And because of this, it is consistent to say that one might observe coordinating behaviour despite the fact that the beliefs the agents have about others contain an internal error. This finding, furthermore, is robust since there is no change in the result even if we admit some variations in the expected composition of a population, or in the depth of reasoning.²⁴ To shed more light on this epistemic aspect of the theory is just the first step in my analysis, but I believe it delivers a non-trivial philosophical finding regarding how decision-making individuals reflect mental states and the reasoning processes of others, and what they might justifiably ascribe to them in coordination by salience. Since correct belief assumption plays a further crucial role, I need a precise notion of it, which may also bring some understanding of what exactly is violated by the cognitive hierarchy model.

Correct belief assumption:

An agent's beliefs—that is, agent's epistemic type (t_i)—for a particular (coordination) interaction are such that she believes that other agent involved in the situation has beliefs (t_j) about her behaviour such that it holds that these beliefs are accurate and correct.²⁵

²³ This condition assumes that the agent has a 'theory of other minds', which holds when he or she is L2 or higher.

²⁴ Imagine a more sophisticated case: a person who is an L3 player and has an expectation that she may interact with an individual of each level with some positive probability (q_0, q_1, q_2). Then we can express her relevant epistemic type for the beverage-choosing game in the following manner:

$$\begin{aligned}
 t_3^{\text{beer}} &: q_2 \times t_{j_2}^{\text{beer}} + q_1 \times t_{j_1}^{\text{beer}} + q_0 \times t_{j_0}^{\text{alcoholic}} \\
 t_{j_2}^{\text{beer}} &\rightarrow t_{i_1}^{\text{beer}} \rightarrow t_{j_0}^{\text{alcoholic}} \\
 t_{j_1}^{\text{beer}} &\rightarrow t_{i_0}^{\text{alcoholic}}
 \end{aligned}$$

Formalization like this allows us to see the profound basis of epistemic asymmetry of cognitive hierarchy theory concerning one strategic aspect (choice of strategy in particular).

²⁵ The correctness simply implies that it is the case that, in the belief hierarchy, agents assume their beliefs about actions and beliefs of their partners are the same

Now this definition and formal description allow us to see one aspect of cognitive hierarchy theory that results from the hierarchical structure and underlying assumptions.²⁶ But we should not lose sight of the fact that *correct belief assumption* expresses a more general idea of coordination and its epistemic context. The case of cognitive hierarchy theory has merely shown at what costs the assumption can be violated if we want to achieve coordination anyway.

How does this analysis help us with variable frame theory? Is there any difference or similarity with respect to the correctness of the agents' beliefs? In section 3, I have briefly explained that variable frame theory proceeds by two distinct steps: structural transformation, and team-coordination. Both make different demands on the individuals involved, yet they are fully adaptable to the formal framework presented and, are, therefore, easy to comprehend and compare. The major difference between the theories lies in the concept of frame that makes salience-based coordination more refined and subtle because it introduces partitions on the strategy set (instead of rather coarse primary salience). If we state that individuals in interaction have the same frame according to which they look upon the coordination problem, then the theory predicts, in this idealized case, that their choices will intersect in a Pareto-optimal result.²⁷ All of this we know already from above, but the question is whether the outcome is in line with correct belief assumption, or against it. We are already familiar with the answer to the question "What would you like me to drink?"—It is water (the single member of *non-alcoholic* group). Nonetheless, it is better to show the epistemic

throughout the nested structure of beliefs (Perea 2012, 145–46). It does not imply that those beliefs are true because there might be many consistent and correct belief combinations, for instance in a game with many Nash Equilibria.

²⁶ To be entirely clear, authors of cognitive hierarchy theory briefly acknowledge this conclusion. (Camerer, Ho, and Chong 2004, 869) My concern relates more to other applications and experiments where this issue is often disregarded.

²⁷ I assume that there is a commonly shared context of interaction which allows the formation of a particular frame. It is not very controversial to proceed in this way because I have already accepted that salience-based coordination involves external factors. Obviously, there are some additional conditions to be fulfilled to secure a clear result (e.g., symmetry disqualification, trade-off principle), for more details see Bacharach (1997).

background of the solution of the game in a similar fashion as before, with only a slight modification in subscripts. Instead of representing cognitive level (which is irrelevant information for variable frame theory), numbers in the agent's epistemic type help us to specify the frame available to him. For instance, t_{i01}^{water} states that both F_0 and F_1 as defined in $\Gamma_{\text{beverages}}$ are families of predicates that agent i (Isaac again) takes into account when he chooses water. Then it is straightforward to express the agents' epistemic types for variable frame theory accordingly:

$$t_{i01}^{\text{water}}: t_{j01}^{\text{water}} \rightarrow t_{i01}^{\text{water}}$$

$$t_{j01}^{\text{water}}: t_{i01}^{\text{water}} \rightarrow t_{j01}^{\text{water}}$$

What we can see immediately is that the epistemic condition of correct beliefs is fully satisfied in this setting, since Isaac (in the first row) expects his partner John will not be mistaken in her beliefs about the Isaac's actions, and vice versa (in the second row). In other words, if Isaac is choosing the beverage with the goal to coordinate, his choice of water is fully justifiable—taking for granted particular frames and team-rationality—by his expectations that John will choose exactly the same, and that he also expects Isaac to choose water.

Variable frame theory predicts that whenever there is a coordination solution induced by salience, agents have beliefs that preserve correctness.²⁸ This result is also consistent for different variations in the structure of a frame. To show a general implication, let me assume a somewhat complex case of a similar game in which one of the agents is aware of an additional family of predicates, say F_x , and he or she recognizes its availability $v(F_x) = p$, where $0 < p < 1$. This describes an aspect of uncertainty, as there is an agent now who may apply some predicates but cannot be sure that the other will do so as well (Bacharach and Stahl 2000, 224). Assume two predicates a and b such that $a \in F_x$, $b \notin F_x$ and for simplicity also that $E(a) \cap$

²⁸ One explanation for these results invokes the well-established theorem that Nash equilibrium in principle rests on correct beliefs, see Tadelis (2013, chap. 5) or Perea (2012, chap. 4). Therefore, variable frame theory as an equilibrium refinement programme carries the same epistemic load.

$E(b) = \emptyset$.²⁹ And strategy s based on expected utility calculation (more specifically based on Bacharach's (1997) trade-off theorem) where $s = \textit{choosing } a$ if $EU(a) > EU(b)$ and $s = \textit{choosing } b$ if $EU(a) < EU(b)$.³⁰ For this reason the epistemic type of the agent i is as follows:

$$t_{i01x}^s: p \times t_{j01x}^s + (1 - p) \times t_{j01}^b$$

$$t_{j01x}^s \rightarrow t_{i01x}^s$$

$$t_{j01}^b \rightarrow t_{i01}^b$$

Even in this general case of framed decision-making, agent i has correct beliefs regarding his co-player's possible choices. Notwithstanding the fact that the agent believes with probability p that the other will be aware of the same frame, and with probability $1 - p$ that j will not notice F_x , the choice of the agent is such that she expects her co-player to think that the same choice is being selected. The result is consistent with correct belief assumption.³¹

Now, I move on briefly to the second epistemic aspect of theories, which is better known and has been already analysed—the issue of *belief in rationality*. The rationality assumption is a cornerstone of decision theory, and from a concise description of both theories, it is intuitively obvious that they depart from the game-theoretical standards. However, I would prefer to show how theories treat the epistemic aspect of rationality, and, thus, to address the question: What does an agent expect regarding the rationality of a co-player? And does the co-player believe in the rationality of the other? It should be evident that I am not dealing with the nature and comparative analysis of rationality requirements which I consider fixed for the respective theories and I have set aside as a separate research agenda.

²⁹ Here I straightforwardly suppose that either choosing a , or choosing b is the team-optimal choice under a condition of the validity of the relevant frame.

³⁰ In the case whereby $EU(a) = EU(b)$, the symmetry disqualification principle rules out both options. Unfortunately, this principle has not been confirmed empirically; see Bacharach and Bernasconi (1997).

³¹ The fact that t_{i01x} might think that the other individual expects her to be t_{i01} does not interfere with the conclusion concerning choices. Moreover, it was my assumption, in this example, to introduce some uncertainty about the other player's frame.

My focus here is fully on the distinction that relates to the epistemic aspects of rationality.

At the start, I need to clarify the epistemic principle I will rely upon in the next analysis.

Belief in rationality:

An agent believes that his co-player is rational, and that the co-player also believes in his rationality.

If we look closely at the scheme of epistemic underpinnings of the solutions to each theory, contrasting features emerge. Variable frame theory explains coordinative behaviour in accordance with the hypothesis that all individuals believe in their co-player's rationality and believe that each partner in coordination believes in their rationality. For instance, epistemic type t_{i01}^{water} , who chooses (team) rational option, expects her co-player to be (team) rational because t_{i01}^{water} think t_{j01} will act in line with the recommendation of the theory: i.e., she will also choose water.³² The result is self-evident from the characterization of her type:

$$t_{i01}^{\text{water}}: t_{j01}^{\text{water}} \rightarrow t_{i01}^{\text{water}}$$

On a more general level of analysis, one may easily see that the theory respects the traditional axiom of the common knowledge of rationality (Tadelis 2013, 64-65). What does it imply about the epistemic background of coordination? A common feature is built into each solution based on variable frame theory implying that players are *epistemically symmetrical* in this important respect. A team-rational player is convinced that she is interacting with someone who also expects the other's actions to be team-rational.

But a contrasting conclusion arises whenever we examine *belief in rationality* in cognitive hierarchy theory. Clearly, it is not that surprising because we recognize that belief hierarchy is based on the gradual nature of rationality. A player at a certain level k is expecting interaction with an individual who is $k-1$ or lower. She also believes that her colleague will

³² Team rationality operates as a criterion for strategy selection in Hi-Lo game, but it is actually belief in rationality that secures equilibrium selection since it is necessary to consider another agent's behaviour and its basis.

expect coordination with someone who is below her actual level, and so on, until L0 is reached. The entire level-based reasoning process is, thus, in principle, grounded in the bounded rationality paradigm, which, as a matter of a fact, means that *belief in rationality* is violated. The following epistemic type scheme captures this result more accurately:

$$t_{i3}^{\text{beer}}: t_{j2}^{\text{beer}} \rightarrow t_{i1}^{\text{beer}} \rightarrow t_{j0}^{\text{alcoholic}}$$

One might raise an objection, pointing to the fact that all types of agents higher than L0 are genuinely rational because they make the best response to their alleged conception of the other's type. However, it is important to bear in mind that *belief in rationality* across different levels is not of the same nature, though all these types are best-responding agents. It is absolutely dissimilar when the agent *i* expects an encounter with t_{j2} or t_{j1} . The first believes that her colleague is rational (L1-rational), while the second does not. After all, it seems that both theories rely on an entirely different epistemic background. In my opinion, the difference could be elegantly depicted in terms of epistemic symmetry and epistemic asymmetry.

5. Epistemic (a)symmetry

The paper raised a thorny question: How can coordination be achieved by salience? I introduced salience as a key factor in establishing a desirable outcome in a coordination game in which communication and precedent are absent, and I presented two recognised explanatory pathways for this phenomenon. The subsequent analysis of the epistemic components of theories has identified remarkable differences regarding correctness of beliefs and belief in rationality. Now, I would like to reveal the last piece in the puzzle by means of which it will be theoretically possible to say that there are two parallel ways of coordinating by salience.

My view of the coordination process via salience respects the fact that the two theories are equally suitable and comprehensible. However, their application is conditioned by specific circumstances affecting the relevant reasoning and decision-making. As we are already well acquainted with the epistemic scaffolding of the theoretical apparatus, I can explain the difference between these two theories by using simple epistemic terms—*epistemic*

symmetry and *epistemic asymmetry*. Also, I argue that they aptly elucidate why both processes of coordination by salience may occur and under what circumstances. Let me explicate these notions as follows:

Epistemic asymmetry in coordination interaction:

In a strategic interaction of a coordination kind, there is epistemic asymmetry if, for every actively participating agent, it holds that he or she does not satisfy correct belief assumption and does not have belief in a co-player's rationality.

Analogically, the notion of epistemic symmetry can be defined in the very same way, except these two essential requirements do hold.

The purpose of these definitions is to build on earlier reflections and to demarcate the relevant context of coordination. In the previous section, I demonstrated several epistemic-type structures as an example of different relationships in the foundation of theories. In light of the above, it is clear that a coordination game with salience allows a number of diverse but parallel procedures. Either I assume that I and my co-player are symmetric in important epistemic aspects, or I expect asymmetric conditions to be valid. In the first situation, correct beliefs and belief in rationality, are prerequisites for the use of subsequent framing and the application of team reasoning. Whereas in the second, the epistemic type of agents is such that they rather anticipate some level of incorrectness in beliefs and uneven standards of rationality, which leads to the utilization of best-response reasoning based on each agent's cognitive efforts.

Imagine we are back in the bar with John and Isaac. How can epistemic conditions affect the resulting coordination? From what I have said, it is quite clear that John and Isaac may end up with the same drink (coordination is achieved) but as a result of different salience-based coordinating mechanism. For instance, If Isaac believes that John considers him to be tired and not caring too much about appropriate reasoning, then Isaac might reliably assess the situation as epistemically asymmetrical. In a sense, "John will think I am tired, and so he will choose beer because he thinks I will just pick something." Moreover, epistemic asymmetry might be even bigger if we assume that John and Isaac are just colleagues from work who rarely meet, and they do not know each other very well. In this scenario,

application of the cognitive hierarchy reasoning looks plausible. On the other hand, by assuming that John and Isaac are close friends who trust each other, we can get an epistemically symmetrical context. John's thoughts might be the following: "Isaac wants us to have the same drink and he knows that I want that too. Isaac will choose water because he thinks that I will be rational, and Isaac will believe that I think water is the best choice for both of us because it is a unique choice of non-alcoholic beverage." Here, it is reliable to say that reasoning described by variable frame theory influence Isaac's and John's decisions. Of course, these two scenarios are just simple stories, but I hope they, at least, illustrate the major point of the paper—the epistemic context matter for coordination by salience.

Another implication of my reflections concerns the relationship between symmetrical and asymmetrical perspective. Based on epistemic grounds, it is more than plausible to say that variable frame and cognitive hierarchy theory are like two sides of the same coin. Under suitable epistemic circumstances, it is likely that an agent is well equipped to use the respective coordinating principle determined by the one theory without ruling out future use of another principle. Thus, I think that this epistemic ramification can help us understand the diversity of coordination procedures and to recognize its contextual value. There might be coordination cases in which individuals consider strategic interaction favourable for epistemic symmetry: they trust their partner or have a positive evaluation of the group, or social bonds are tight and firm, etc. (Bacharach 2006, chap. 3; Colman, Pulford, and Rose 2008; Hindriks 2012). For all these factors influencing our perception of the epistemic environment of coordination games it seems legitimate to predict the outcome in accordance with variable frame theory. On the other hand, some conditions—payoff bias (Crawford, Gneezy, and Rottenstreich 2008) or prudential thinking (Cooper et al. 1990)—seem to fit with epistemic asymmetry, and favour cognitive hierarchy theory.

Finally, what does my conclusion say about the impact on experimental research? In most situations these theories imply the same result, yet when it comes to test cases predictions may diverge in various ways. My claim is that although we observe similar results (Mehta, Starmer, and Sugden 1994; Bardsley et al. 2010; Colman, Pulford, and Lawrence 2014) it proves little,

as we can still defend both theories. In my view, coordination by salience is a result of the involvement of two processes whose active impact on the final outcome is fundamentally—but not solely—determined by the epistemic niche of a given interaction. The challenge for future research might be to test experimentally factors entrenched in the epistemic conditions which interfere with coordinating decisions and cause behavioural variations.

6. Conclusion

In this paper, I have shown that if one wants to properly understand coordination by salience, it seems necessary to take into account the epistemic restrictions that are imposed on reasoning in different coordination procedures. Consequently, the two well-known and prominent theories, variable frame theory and cognitive hierarchy theory should be regarded as complementary ways of explanation of salience-based coordination. Besides, I have suggested that the criteria of epistemic symmetry and epistemic asymmetry comprehensively specify tacit assumptions of theories and shed a light on the important difference build into their foundations. In variable frame theory, correct beliefs and belief in rationality, are prerequisites for the use of subsequent framing and the application of team reasoning. Whereas in cognitive hierarchy theory, the epistemic type of agents is such that they rather anticipate some level of incorrectness in beliefs and uneven standards of rationality, which leads to the best-response reasoning based on epistemic asymmetry.

References

- Bacharach, Michael. 2006. *Beyond Individual Choice: Teams and Frames in Game Theory*. Edited by Natalie Gold and Robert Sugden. Princeton: Princeton University Press. <https://doi.org/10.1515/9780691186313>
- Bacharach, Michael, and Michele Bernasconi. 1997. “The Variable Frame Theory of Focal Points: An Experimental Study.” *Games and Economic Behavior* 19 (1): 1–45. <https://doi.org/10.1006/game.1997.0546>
- Bacharach, Michael, and Dale O. Stahl. 2000. “Variable-Frame Level-N Theory.” *Games and Economic Behavior* 32 (2): 220–46. <https://doi.org/10.1006/game.2000.0796>

- Bardsley, Nicholas, Judith Mehta, Chris Starmer, and Robert Sugden. 2010. "Explaining Focal Points: Cognitive Hierarchy Theory versus Team Reasoning." *Economic Journal* 120 (543): 40–79. <https://doi.org/10.1111/j.1468-0297.2009.02304.x>
- Camerer, Colin, Teck-Hua Ho, and Juin-Kuan Chong. 2004. "A Cognitive Hierarchy Model of Games." *The Quarterly Journal of Economics* 119 (3): 861–98. <https://doi.org/10.1162/0033553041502225>
- Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.
- Colman, Andrew M., Briony D. Pulford, and Jo Rose. 2008. "Collective Rationality in Interactive Decisions: Evidence for Team Reasoning." *Acta Psychologica* 128 (2): 387–97. <https://doi.org/10.1016/j.actpsy.2007.08.003>
- Colman, Andrew M., Briony D Pulford, and Catherine L Lawrence. 2014. "Explaining Strategic Coordination: Cognitive Hierarchy Theory, Strong Stackelberg Reasoning, and Team Reasoning." *Decision* 1 (1): 1–36. <https://doi.org/10.1037/dec0000001>
- Cooper, Russell W., Douglas V. Dejong, Robert Forsythe, and Thomas W. Ross. 1990. "Selection Criteria in Coordination Games: Some Experimental Results." *The American Economic Review* 80 (1): 218–33.
- Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich. 2008. "The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures." *American Economic Review* 98 (4): 1443–58. <https://doi.org/10.1257/aer.98.4.1443>
- de Bruin, Boudewijn. 2009. "Overmathematisation in Game Theory: Pitting the Nash Equilibrium Refinement Programme against the Epistemic Programme." *Studies in History and Philosophy of Science Part A* 40 (3): 290–300. <https://doi.org/10.1016/j.shpsa.2009.06.005>
- Geanakoplos, John. 1992. "Common Knowledge." *Journal of Economic Perspectives* 6 (4): 53–82. <https://doi.org/10.1257/jep.6.4.53>
- Gilbert, Margaret. 1989. "Rationality and Salience." *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 57 (1): 61–77. <https://doi.org/10.1007/bf00355662>
- Gold, Natalie, and Robert Sugden. 2007. "Collective intentions and team agency." *Journal of Philosophy* 104 (3): 109–37. <https://doi.org/10.5840/jphil2007104328>
- Gold, Natalie, and Jurgis Karpus. 2016. "Team Reasoning: Theory and Evidence." In *Routledge Handbook of Philosophy of the Social Mind*, edited by Julian Kiverstein. Routledge-Taylor Francis. <https://doi.org/10.4324/9781315530178>
- Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press.

- Haruvy, Ernan, and Dale O. Stahl. 2007. "Equilibrium Selection and Bounded Rationality in Symmetric Normal-Form Games." *Journal of Economic Behavior and Organization* 62 (1): 98–119. <https://doi.org/10.1016/j.jebo.2005.05.002>
- Hindriks, Frank. 2012. "Team Reasoning and Group Identification." *Rationality and Society* 24 (2): 198–220. <https://doi.org/10.1177/1043463111429274>
- Janssen, Maarten. 2001. "On the Principle of Coordination." *Economics and Philosophy* 17 (2): 221–34. <https://doi.org/10.1017/s0266267101000244>
- Lewis, David. 1969. *Convention. A Philosophical Study*. Harvard University Press. <https://doi.org/10.1002/9780470693711>
- Luce, Duncan, and Howard Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey*. New York: Dover Publications.
- Mehta, Judith, Chris Starmer, and Robert Sugden. 1994. "The Nature of Salience: An Experimental Investigation of Pure Coordination Games." *American Economic Review* 84 (3): 658–73.
- Ohtsubo, Yohsuke, and Amnon Rapoport. 2006. "Depth of Reasoning in Strategic Form Games." *Journal of Socio-Economics* 35 (1): 31–47. <https://doi.org/10.1016/j.socec.2005.12.003>
- Perea, Andrés. 2012. *Epistemic Game Theory. Reasoning and Choice*. New York: Cambridge University Press. <https://doi.org/10.1017/CBO9780511844072>
- Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
- Sillari, Giacomo. 2008. "Common Knowledge and Convention." *Topoi* 27 (1): 29–39. <https://doi.org/10.1007/s11245-008-9030-7>
- Stahl, Dale O., and Paul W. Wilson. 1995. "On Players' Models of Other Players: Theory and Experimental Evidence." *Games and Economic Behavior* 10 (1): 218–54. <https://doi.org/10.1006/game.1995.1031>
- Sugden, Robert. 1995. "A Theory of Focal Points." *The Economic Journal* 105 (430): 533–50. <https://doi.org/10.2307/2235016>
- Sugden, Robert. 2011. "Salience, Inductive Reasoning and the Emergence of Conventions." *Journal of Economic Behavior and Organization* 79 (1–2). <https://doi.org/10.1016/j.jebo.2011.01.026>
- Tadelis, Steven. 2013. *Game Theory: An Introduction*. Princeton: Princeton University Press.
- Ullmann-Margalit, Edna. 1977 (2015). *The Emergence of Norms*. Oxford: Oxford University Press.
- Young, H Peyton. 1996. "The Economics of Convention." *The Journal of Economic Perspectives* 10 (2): 105–22. <https://doi.org/10.1257/jep.10.2.105>

Why Did You Really Do It? Human Reasoning and Reasons for Action

José Ángel Gascón*


Received: 26 July 2020 / Revised: November 1 2020 / Accepted: 23 November 2020

Abstract. During the last decades several studies in cognitive psychology have shown that many of our actions do not depend on the reasons that we adduce afterwards, when we have to account for them. Our decisions seem to be often influenced by normatively or explanatorily irrelevant features of the environment of which we are not aware, and the reasons we offer for those decisions are a posteriori rationalisations. But exactly what reasons has the psychological research uncovered? In philosophy, a distinction has been commonly made between normative and motivating reasons: normative reasons make an action right, whereas motivating reasons explain our behaviour. Recently, Maria Alvarez has argued that, apart from normative (or justifying) reasons, we should further distinguish between motivating and explanatory reasons. We have, then, three kinds of reasons, and it is not clear which of them have been revealed as the real reasons for our actions by the psychological research. The answer we give to this question will have important implications both for the validity of our classifications of reasons and for our understanding of human action.

Keywords: Cognitive psychology; explanation; justification; motivation; rationalization; reasons for action.

* Universidad Católica del Maule

 <https://orcid.org/0000-0001-5571-6602>

 Department of Philosophy, Universidad Católica del Maule. Avenida San Miguel 3605, Talca, Chile

 jgascon@ucm.cl



1. Introduction

Human beings consider, at least sometimes, what reasons we have to do something. When we do, according to a widespread view, what happens is the following: we usually act in the light of those reasons that seem to us to be the best, and then justify our action before others by putting forward the reasons that moved us to act. Imagine, for example, that I have been offered two jobs, one of which has a better salary and the other is in a city that I like. Which job I will accept depends on my weighing of those reasons (and others) and of which considerations are more important to me. After my decision, if I am challenged to justify it—“why did you do that?”—I will present those reasons that made me opt for one job rather than the other, hoping that my audience will see why I made the best choice. I will, then, engage in argumentation in order to show that those reasons that moved me to act in a certain way were the best reasons, all things considered.

This can be considered as the standard, common-sense view of reasoned action and justification of actions. We justify our actions by presenting reasons, and those are precisely the reasons for which we acted. Characterisations of the rational agent or the critical thinker which focus on reasons—as opposed to those which focus on the suitability of means to an end, for instance—tend to rely on this view of the relationship between reasons and action. According to Harvey Siegel, for example, a critical thinker is someone who is “appropriately moved by reasons” (1997, 49). And, in the literature on normativity and practical reasons, authors such as Scanlon (2014) and Kiesewetter (2017) define rationality in terms of responsiveness to reasons.

The idea that we should justify our actions by putting forward the reasons for which we acted seems like a plausible one. After all, there is a tendency to see people as irrational—or, at the very least, hypocritical—when they act for one reason and afterwards attempt to justify their action by appealing to different reasons. Consider the case of someone who decides to study philosophy and holds that her reason for that decision is her love of knowledge, when in fact what moved her towards a philosophical career is her desire to enjoy a high cultural status. No doubt many of us would see

that behaviour as falling short of rationality—or, if she is aware of her real motivation, as insincere.

However, if this is how we should understand the rational justification of actions, then apparently we are in serious trouble. The empirical research in the psychology of reasoning has shown that human beings are very bad at identifying the causes of *our own* actions. A growing number of empirical studies have provided evidence that we lack access to the knowledge of what considerations move us when we act. The reasons that we put forward when we are challenged to justify our behaviour are not, it seems, those reasons *for which* we acted, but merely our best guesses about why we acted that way—even though no doubt those guesses are sometimes right.

How worried should we be by this conclusion? The purpose of this article is to give a tentative answer to this question. I believe that any such answer, if it is to be plausible, must be both philosophically and psychologically informed. Our philosophical accounts of practical reasoning need to take into account the empirical findings that indicate what feats human reason can and cannot achieve; and, at the same time, the psychological research must be based on a philosophical understanding of reasons so that it is clear what conclusions can and cannot be drawn from the empirical data. I will begin, in the next section, by reviewing the empirical studies in psychology of reasoning that cast doubt on our ability to detect the reasons that move us to act. Then, in section 3, I will present philosophical distinctions between kinds of reasons and I will provide an interpretation of the conclusions of psychological studies in the light of those distinctions. Finally, in section 4, I will draw some preliminary conclusions about how all this should affect our conceptions of justification of actions and of rationalisation.

2. Psychological research on reasons for action

Up until the 1970s, it was widely assumed by psychological researchers that we are aware of the mental processes that lead to our judgements and our behaviour (Kunda 1999, 265). In order to study people's choices and evaluations, investigators resorted to self-report questionnaires in which the participants in the experiments were asked to state why they behaved as they did. Researchers who attempted to study the grounds for voting for a

political candidate or for choosing a job, for example, simply asked people why they voted for a certain candidate or why they chose a certain job. However, it eventually became manifest that such self-reports are not reliable.

In their ground-breaking article, Nisbett and Wilson (1977) reviewed a series of empirical studies in which a particular stimulus demonstrably influenced the participants' actions and judgements but, when interviewed, the participants denied that influence and tended to explain their behaviour by reference to other factors. An example is the large number of experiments that showed the existence of the "bystander effect," the fact that people are less likely to help a person in distress if there are many other onlookers around (Latané and Darley 1970). After the experiments, Latané and Darley asked the participants whether their decision to help or to abstain from helping had been influenced by the presence of other people. Despite the robust evidence that showed that a greater number of onlookers correlated with a failure to help, the participants systematically denied that influence. As the authors explain (*Ibid.*, 124):

We asked this question every way we knew how: subtly, directly, tactfully, bluntly. Always we got the same answer. Subjects persistently claimed that their behavior was not influenced by the other people present.

Nisbett and Wilson also conducted a series of small studies in order to investigate the accuracy of causal explanations of one's own behaviour (Nisbett and Wilson 1977; Wilson and Nisbett 1978). The experiments were designed in a way that resembled as close as possible situations of the real life, with little or no deception involved. Yet they were also designed so that the stimuli that would probably influence the participants' behaviour were of a counter-intuitive sort and hence their influence could not be accounted for by the participants' prior causal theories of how people behave (Nisbett and Wilson 1977, 242). Therefore, those stimuli could only have been detected by the participants if they had genuine introspective access to their own cognitive processes. As expected, people were influenced by factors whose influence they could not detect—and, interestingly, the researchers themselves were highly unsuccessful in their predictions of which factors would influence them.

In one of those studies (Ibid., 243), the participants were asked to evaluate four pairs of stockings. They had to choose one of those pairs and, afterwards, they were asked why they had chosen it. The trick was that all the stockings were identical. Nisbett and Wilson observed that the stockings situated towards the right were preferred over the ones situated at the left. However, when the participants were asked about the reasons for their choices, the position of the article was never mentioned. In fact, when the researchers suggested that possibility to the participants, they denied it. The authors explain that (Wilson and Nisbett 1978, 124):

Only a quarter of the subjects required any prompting to explain the basis of their choices. Most of the subjects promptly responded that it was the knit, weave, sheerness, elasticity, or workmanship that they felt to be superior. [...] Not a single subject mentioned the position of the stockings as a reason for the choice.

Not only do we often fail to detect factors that cause our behaviours, but we also tend to report as reasons for our choices and judgements stimuli that actually had no effect on us. For example, in another experiment (Nisbett and Wilson 1977, 246), the participants had to predict how much electric shock they would take. Some of them were said that the shocks would do “no permanent damage,” while the others were not given that reassurance. Then, the researchers asked the first group whether that comment had affected their predictions, and they asked the second group whether, had they made that comment, their predictions would have been different. Inclusion of the reassurance proved to have no effect on the predictions of how much shock the participants would take, but a majority of them reported that it affected their predictions.

What all this evidence shows is not merely that we are sometimes wrong when we report our reasons for our decisions and judgements—that would hardly be big news. Neither can it be concluded that we are *always* wrong; as Nisbett and Ross (1980, 211) admit, we are often accurate in our explanations of the reasons for our behaviour. The worrying implication of that research on self-reports is rather that we *lack introspective access* to the reasons that guide our behaviour. The process by which we arrive at a belief of why we did something is the same whether that belief is accurate or inaccurate: we *infer* it from the known data and from our prior theories of

human behaviour. That is, it is the same process that we follow when we propose causal explanations of *other people's* behaviour (Ibid.). If, for example, I buy a bottle of water and I claim that I did so because I was thirsty, I am surely right. But this is so simply because we have a common-sense theory of why people usually buy bottles of water, and that theory is largely correct. Notice, also, that it would be just as easy to identify the reason why someone else bought a bottle of water. This was Nisbett and Ross's conclusion (1980, 211):

Empirically, this means that under most circumstances subjects will be right in their causal accounts if and only if observers, working with similar externally available information, also are right.

The problems begin when there is no prior theory or when that theory does not fit the case at hand. If we fail to help a person in distress because there are many other people around, or if we choose a pair of stockings because they are situated on the right, then we are likely to give a wrong account of our behaviour, since we have no prior theory about the relationship between those reasons and those actions. And, in those cases, we are just as likely to be wrong about our own behaviour as we are to be wrong about other people's behaviour. The process is the same in both cases.

Of course, when it comes to our own behaviour we have access to data that we lack when we attempt to interpret someone else's behaviour, such as our feelings, explicit goals, beliefs or memories (Nisbett and Ross 1980, 203). However, Wilson argues that this private information can also mislead us. He points out that "the vast amount of inside knowledge we have about ourselves increases confidence in our self-knowledge, but does not always lead to greater accuracy" (Wilson 2002, 113). A stranger with no access to that information could be more accurate about the causes of our actions, and in fact this seems to be often the case. He concludes (Ibid., 112):

Averaging across several studies, there seems to be no net advantage to having privileged information about ourselves: the amount of accuracy obtained by people about the causes of their responses is nearly identical with the amount of accuracy obtained by strangers.

If that conclusion is correct, then many of the processes that cause our actions and judgements are unconscious, just as the processes that are responsible for perception or textual comprehension. Kunda (1999, 270ff.) reviews other studies that provide evidence of those unconscious processes that influence our behaviour, including aspects such as implicit memory and subliminal perception. Some cognitive scientists have accepted the most dramatic implications of this conclusion regarding our conscious will. Evans, while admitting that there is a difference between voluntary and involuntary actions, questions the very existence of a conscious will (2010, 177):

‘We’ are not conscious persons in control of our behaviour and the reflective mind does not equal a conscious mind. The conscious person is a construction of the brain, an illusory narrative that accompanies us through life.

In the same vein, Wegner (2002) talks about the conscious will as an “illusion.” According to his theory of apparent mental causation, conscious will is not a cause of actions but simply a (possibly misguided) *feeling* that an action was caused by us. He explains (Ibid., 336):

Apparent mental causation suggests that the experience of consciously willing an act is merely a humble estimate of the causal efficacy of the person’s thoughts in producing the action. Conscious will is the mind’s way of signaling that it might have been involved in causing the action. The person’s experience of doing the act is only one source of evidence regarding the actual force of the person’s will in causing the action, however, and it may not even be the best source.

Although Wilson does not endorse the conclusion that conscious will is *always* an illusion, he admits that very often it is (2002, 48): “We may have the impression that we, our conscious selves, are in complete control, but that is at least in part an illusion.”

Now, if we accept these psychologists’ conclusion that we tend not to be (or perhaps never are) introspectively aware of what factors influence our actions and judgements, and in fact we are often wrong about them, the question is: how big a problem is that for our philosophical theories about

reasons for action and justification of actions? This is a very broad issue that cannot be solved in a single paper. As a first step, however, it would help to be clear about what exactly Nisbett and Wilson's experiments uncovered. Did they identify our *real* reasons for action? Or did they show us simply the *causes* of our actions? Are they the same thing? Sorting out this conceptual issue is the purpose of this paper, and to this I move in the next section.

3. What reasons are we talking about?

The results of the experiments conducted by Nisbett and Wilson certainly seem to reveal something important that jeopardises our ideas of intentional action and justification of actions. But, what is it exactly that was identified in those experiments? In their articles, Nisbett and Wilson used a variety of terms to refer to the stimuli that influenced the participants' behaviour: "influences," "explanation," "causes," "causal factors," and "reasons" for choice. The point was that there seemed to be a mismatch between the reasons stated by the participants in the studies and whatever it was—influences, reasons, causes—that explained their choices. Thus, a necessary first step in the assessment of the implications of those studies for our philosophical theories is the clarification of these factors that explained the participants' behaviour.

The most natural interpretation, I believe, is that the experiments identified the *causes* of our actions and judgements. Now, it is well known that, according to some philosophical views, reasons for action just are the causes of our actions. Davidson (1963) famously argued for that view. If that is how we should understand practical reasons, then the discovery that people lack direct awareness to the causes of their actions obviously challenges our practice of giving reasons for our actions. If we cannot detect the causes of our behaviour, and practical reasons are precisely those causes, then it seems that the reasons we give for our actions are mere speculations. In that case, we cannot be sure for what reasons we did something, just as we cannot be sure how our stomach is digesting what we ate.

However, the philosophical literature has distinguished between different kinds of reasons, and Davidson focused on only one of them: the kind

of reason that “explains the action by giving the agent’s reason for doing what he did” (Ibid., 685). Beyond reasons that explain the motivations of agents, there are also reasons that justify their actions. It is one thing to report what considerations *motivated* us to do something, and hence explain our action; it is something different to *justify* our actions with considerations that make them the right thing to do (Dancy 2000, 20–25). The former kind of reasons has been called *motivating* reasons, whereas the latter has been called *normative* reasons. Thus, Parfit (1997, 99) says that normative reasons are those that we are looking for when we ask “What do we have most reason to want, and do?”; motivating reasons, on the other hand, are those in light of which we act. Dancy explains the distinction this way (2000, 2):

There is the question what were the considerations in the light of which, or despite which, he acted as he did. This issue about *his reasons for doing it* is a matter of motivation. There is also the question whether there was good reason to act in that way, as we say, *any reason for doing it* at all, one perhaps that made it sensible in the circumstances, morally required, or in some other way to be recommended, or whether there was more reason not to do it. [...] This second question raises a normative issue.

We can act for a good reason, in which case our motivating reason is also our normative reason, but it is also possible for these two kinds of reasons to diverge. Imagine, for example, that I voted for a certain political candidate because it seemed to me that she was the most honest and competent one. Those were the reasons that I considered when I was deciding my vote, so those are the reasons for which I acted. When asked, I offer those reasons to justify my choice. In this case, my normative reasons are the same as my motivating reasons. But let us imagine a slightly different scenario. Imagine that, even though that political candidate was indeed the most honest and competent one, I did not take that fact into account when deciding my vote; instead, what motivated me to vote for her was that she was born in the same city as me. I still justify my vote before others by mentioning her honesty and competence, but I know that I voted for her because we were born in the same place. In this second case, my motivating reasons are different from my normative reasons.

Let us differentiate, then, between:

Normative reasons: Considerations that make an action the right thing to do, that count in favour of doing that action.

Motivating reasons: Considerations that moved me to do something, those in the light of which I acted.

As Dancy (2000, 2) and Alvarez (2009) argue, this reference to two “kinds” of reasons should not be understood as implying that there are really two sorts of reasons—reasons that motivate and reasons that justify. They are different kinds of reasons only in the sense that they are offered in answer to two different questions: (1) what makes that action right?, and (2) why did you do that action?

Now, if we go back to Nisbett and Wilson’s experiments and ask what kind of reasons—if any—they have discovered, it seems clear that we can rule out normative reasons. The researchers, as we have seen, deliberately designed the experiments so that there were no good reasons to prefer one pair of stockings rather than another—they were all of the same quality. The reasons that the participants gave—the superior knit, weave, sheerness, elasticity, or workmanship—were false, and it is a commonplace in philosophy that bad normative reasons are not reasons at all¹. If anything, Nisbett and Wilson showed that the participants in the experiments could not offer any normative reasons. The experiments certainly did not uncover any normative reasons for there was none in those cases.

Yet, it is not obvious to me either that the findings of the experiments refer to motivating reasons. Those findings do refer to factors that explain people’s actions, but motivating reasons are not simply any kind of explanation; motivating reasons explain actions only insofar as those actions were made *in the light of* reasons. That means that a causal factor would not count as a motivating reason if the agent has not consciously considered it and decided to act on the basis of it. For a cause of people’s actions to be

¹ Dancy puts it at the very beginning of his book (2000, 1-2): “A bad reason for doing something, if it is not merely a not very good reason for doing it, can only be no reason at all for doing it; if so, it is not a reason in the sense intended, since it does not favour the relevant action.”

a motivating reason, they must at least recognise it as a reason and be guided by it. Suppose, for instance, that I fell on the street because a car hit me. It does not make sense to say that my motivating reason for falling was that a car hit me, i.e. that I was motivated to fall by the hit of a car. No doubt the hitting of the car explains the event, but it is not an explanation in terms of *motivation*.

There is, however, a third possibility besides normative and motivating reasons: what Searle (2001, 111) calls “straight causal explanations” and Dancy (2000, 5) calls “reasons why.” These reasons do not involve considerations that the agent takes to favour some action. Actions that are explained by “reasons why” are not performed in light of those reasons, but simply caused by them. This is the case with the explanation of the fact that I fell on the street that mentions the hit of the car. Many other events involve this kind of explanations, in which no reason was considered by the agent, as Dancy reminds us (*Ibid.*):

What explains why one person yawned may be that someone else yawned just next to them. What explains why he responded so aggressively may be that he is having trouble at home or that he has taken a particular form of medication. What explains why he gave this student a better grade than she deserved is that he was unconsciously influenced by the fact that she always dresses so neatly (or something even less defensible). What explains why so many people buy expensive perfume at Christmas is the barrage of advertising on the television. What explains why he didn't come to the party is that he is shy. In none of these cases are we specifying considerations in the light of which these things were done.

Dancy states that what these explanations involve “is not a reason at all, really, but rather a cause” (*Ibid.*, 6). However, Alvarez (2009, 184) argues that its being a cause does not exclude its being a reason, since both terms belong to different domains: that of causation and that of explanation. We use reasons to explain actions, and those reasons sometimes happen to be causes in the natural realm. Therefore, she proposes that, besides normative and motivating reasons, we should consider *explanatory* reasons. If we differentiate among different kinds of reasons on the basis of the role

they play in answering different questions, then the question of what explains an action is substantially different from the question of what motivated the agent—even though, of course, the same reason can answer both questions.

We have, then, according to Alvarez's proposal, three kinds of reasons:

Normative reasons: Considerations that make an action the right thing to do, that count in favour of doing that action.

Motivating reasons: Considerations that moved me to do something, those in the light of which I acted.

Explanatory reasons: Considerations that explain why I did something, what caused my action.

Explanatory reasons are a better candidate for the kind of reasons that the psychological experiments revealed. They are causes that explain the participants' actions without being at the same time motivating reasons, since the participants did not consider them and even denied their influence. Those reasons are causes in the same sense that taking a certain medication is the cause of aggressive behaviour. They influence our actions but we are unaware of that influence.

There is one crucial difference between explanatory reasons and the other two kinds of reasons, and that difference is what makes the findings of psychological experiments so shocking: explanatory reasons do not necessarily involve *human agency*. Just as they can be used to explain human actions, they are also what explains events such as the rain, the collapse of a building or the movement of waves at sea. There are no normative or motivating reasons for events like these—water and buildings do not consider reasons and do not attempt to justify their actions. So, when human actions are explained on the basis of explanatory reasons that are not also normative or motivating, that certainly feels like our sense of agency itself is being challenged. That may be all right for certain human actions, such as yawns or sudden outbursts of aggressiveness, but it is frightening to find out that it also involves actions for which we believe we have motivating reasons, such as choosing stockings of helping a person in distress. No wonder some cognitive scientists have concluded that conscious will is an illusion.

Should we then give up any talk of reasons altogether? Even though satisfactorily solving this issue would require a longer discussion, and the main topic of the present article was to sort out the kinds of reasons that the psychological experiments are referring to, in the next section I will outline a path that—in my view—we should take. In order to sketch an answer that question, I believe we must go beyond Nisbett and Wilson's experiments and consider the role of a kind of reasons that initially did not seem empirically relevant: normative reasons.

4. Justification, motivation and rationalisation

Normative reasons are importantly different from explanatory and motivating reasons. The question of what considerations count in favour of an action, what considerations make an action right, is not empirical but normative. The psychological research can test our conceptions of explanatory and motivating reasons—of what explains our actions and what motivates us—but only the philosophical reflection can test our views on normative reasons. What is right is right regardless of whether it explains our actions or motivates us. That is why Nisbett and Wilson's experiments could not throw any light on normative reasons.

However, what interests us here is not merely whether normative reasons exist. For normative reasons to be something more than a philosophical construct, they must influence our actions somehow. If the reasons which justify our actions have no influence on our behaviour, as the experiments that we have seen suggest, then it begins to look as if those reasons were merely *epiphenomenal*: they would play no role in the determination of our actions.

How could we measure the causal efficacy of normative reasons? The safest way, I believe, is to focus on the reasons that we offer with the deliberate purpose of justifying actions. Even though we often attempt to justify our actions by explaining why we performed them—i.e. by citing motivating reasons—normative reasons need not be also motivating reasons (Dancy 2000, 3). Sometimes we simply argue that what we did was right without intending to explain what moved us to do it. Someone might, for example, argue that the choice of her academic career was a good one—because, say,

it has good job prospects and it fits her character—without even remembering why she chose it in the first place. When we attempt to justify an action this way, what we offer is *purported* normative reasons—these include both genuine normative reasons and bad reasons, which, as was pointed out in the previous section, cannot be considered reasons. Sometimes the reasons with which people attempt to justify their actions are good ones, and sometimes they are false and hence they are not really reasons. Normative reasons, therefore, are a *subset* of the purported normative reasons that people can offer.

Thus, my main claim in this section is the following: if the reasons that we consider and offer to justify actions play a causal role in our behaviour, then it follows that normative reasons are causally efficacious. That is, if purported normative reasons influence our behaviour, and normative reasons are a subset of purported normative reasons, then normative reasons must influence our behaviour. In plain words, if our actions are influenced by reasons which we think (correctly or incorrectly) that would justify our actions, then it can be said that at least sometimes our actions are influenced by reasons that do justify our actions. It would be very odd indeed if we were influenced by reasons but only the bad ones.

In fact, we have evidence that shows that at least sometimes purported normative reasons motivate our decisions and beliefs. The idea that people can take decisions and change their minds on the basis of reasons that show that some action is the right thing to do seems to be a necessary assumption in order to account for much of human behaviour. This can be seen most clearly in psychological experiments involving interpersonal argumentation. As Mercier and Sperber (2017, 264–265) point out, groups of people are more able to solve logical problems than individuals working alone, and this happens because people working in groups benefit from the exchange of reasons. For example, Trouche, Sander and Mercier (2014) showed that people who are confronted with arguments or who argue are more likely to solve logical problems such as those of the Cognitive Reflection Test (Frederick 2005) and others. Their experiments were designed in a way that ruled out the effect of degrees of confidence of some participants on others, measuring specifically the effects of good argumentation. Thus, they concluded that their results “make it clear that arguments, rather than confidence, are

the main factor explaining the performance of groups discussing intellectual tasks” (Ibid., 1968).

According to Mercier and Sperber, the main function of the human faculty of reason is not to make better decisions but—at least sometimes—to make decisions for which we can come up with (allegedly) good reasons (Mercier and Sperber 2017, 255): “when people have weak or conflicting intuitions, reason drives them toward the decision for which it is easiest to find reasons—the decisions that they can best justify.” According to their argumentative theory of reasoning, the justifications that we offer or that we mentally rehearse do guide our actions. If purported normative reasons are understood as attempts to justify actions—as I have assumed here—then they seem to influence decisions. Purported normative reasons can influence us in group discussion, as Trouche et al. showed; or, even when there is no interpersonal argumentation taking place, the prospective justification that we mentally rehearse leads us in the direction of the most acceptable reasons.

It may be thought that this conclusion clashes with certain experiments that show that rational argumentation rarely changes people’s minds, particularly in the moral realm (Haidt 2001). I admit that sometimes that may be the case and that the power of normative reasons is somehow limited. Nevertheless, this does not mean that they never have any effect. Cohen (2019, 715) addresses this problem and notes that:

The point is that the marginalization of reason as only *rarely* effective is also an acknowledgement that it sometimes is effective. The claim is not that there is no causal footprint for reasoning and argumentation at all; rather, the claim is that the effects are limited.

In fact, just as there is evidence that arguments often fail to convince people in certain domains, we have also evidence that sometimes—not infrequently, I would say—arguments change people’s minds. As we have seen, the experiments that Mercier and his collaborators carried out and reviewed show exactly that. Moreover, I would like to add the observation that, even when the participants in an experiment fail to be convinced by arguments when they should, all the experiments have shown is that people have not been convinced *immediately*. There is still the possibility that

people keep thinking about the reasons they have heard and change their minds *in the long run*. In fact, Kjeldsen (2020) interviewed people about their experiences of changing their minds on important social or political issues and he found out that it took them between 4 and 9 years to do so.

Hence, normative reasons do not seem to be inert. They can lead people to take a decision or form a judgement, even if it takes time. But then, when such a thing happens, we can confidently say that those are people's *motivating* reasons. Just as in Nisbett and Wilson's experiments there was no way that the motivating reasons reported by the participants played any role in determining behaviour, in the experiments reviewed by Mercier and Sperber there seems to be no alternative to granting normative reasons a causal role. Therefore, reasons are not a mere epiphenomenon; motivating reasons, i.e. conscious reasons with a causal power on our decisions, exist.

Concluding that motivating reasons exist, however, is not much if those reasons can only be reliably detected in the laboratory. And here lies precisely the lesson that we should draw from the psychological research: we do not have direct access to the causes of our own actions, we just infer the possible causes from a body of data and a more or less accurate theory of human behaviour, so we can always be wrong about our alleged motivating reasons. We should not be confident that we did something for the reasons that we think we did it. We need to accept our unreliability even in the realm of our own actions. As Wilson suggests (2002, 113): "we all might want to be more humble about the accuracy of our causal judgments."

So one lesson regarding our own self-reports is that we should acknowledge the possibility that *we* are wrong. What about other people's accounts? In my view, taking that conclusion seriously should lead us to giving considerably less weight to motivating reasons in people's attempts to justify their actions. Accounts of why people did something should not be given a predominant place in justifications of actions. When it comes to justification, we should focus on normative reasons, and these should be kept separate from motivating reasons. I believe this is a conclusion that we should accept in light of the unreliability of our reports of motivating reasons. If we do not want the weakness of those reports to be transferred to our practice of justifying actions, the kind of reasons that make an action

right or wrong should be relatively independent of the kind of reasons that explain why that action was done.

This may seem too radical a proposal, as it risks blurring the distinction between a genuine justification of an action and a *rationalisation*. Without that distinction, it may be thought that the very idea of rationality is in danger. I will use the rest of this section to try to dispel that worry.

In the most common sense of the term, a rationalisation is a purported account offered by an agent of one of her actions that (Audi 1985, 163):

Offers one or more reasons for doing that action.

Represents his doing that action as at least *prima facie* rational given those reasons.

Does not explain why the agent did that action.

That is, a rationalisation is an attempt to *justify* an action (point 2) by offering *normative* reasons (point 1) that are not at the same time *motivating or explanatory* reasons (point 3). Someone might, for example, justify his decision not to eat peppers by asserting that they are bad for his health, when in fact his motivating reason is simply that he does not like them. But rationalisation so defined is exactly what, I have argued, theories of rational action should be more tolerant of. Should we then accept the fact that, as Mercier and Sperber (2017, 253) say, humans are rationalisation machines? In that case, it seems that we would be condoning widespread irrationality. The problem with the empirical studies that show that rationalising is what the human mind usually does is that they seem to warrant the conclusion that we are all irrational. As Cohen (Cohen 2019, 711) puts it:

[...] a great deal, perhaps even most, of our reasoning turns out to be *rationalizing*. The reasons we give for our positions are seldom either the real motives or the effective causes of why we have those positions. The uncomfortable conclusion, unfortunately substantiated by too many empirical studies to dismiss, is that we are not as rational as we like to think.

However, I believe that we can dissipate (at least most of) the fear of irrationality if we do not underestimate the extent to which normative

reasons can be criticised. Justifications, even if they are rationalisations, can still be correct or incorrect. A match between purported normative reasons and motivating reasons is not the only way to check the correctness of justifications—it is not, in fact, the main one or even the most demanding one. Purported normative reasons by themselves must fulfil several criteria for them to constitute a satisfactory justification. One of those criteria is, of course, that they must be true, they must mention real facts. This criterion alone allows us to see where the participants in Nisbett and Wilson's stockings experiment got their justification wrong: they mentioned particular features that allegedly made certain stockings the best ones in the lot, whereas in fact all of them were identical. There is no need to appeal to their motivating reasons in order to conclude that their justifications were flawed.

Besides truth, we should also expect an agent's purported normative reasons to *cohere* with those that the same agent has offered in similar circumstances. The principle that like cases should be treated alike is firmly established both in law and in ethics, but it is also relevant in other domains. This principle helps us explain what might be wrong in the justifications offered by the participants in the bystander effect experiments performed by Latané and Darley. Surely, those who helped the person in distress when there were few onlookers could have justified their action by saying that the person needed help, but if they do not help in a similar scenario with more onlookers, there might be an incoherence in the normative reasons they state.² Again, as in Nisbett and Wilson's experiments, we can talk about the rationality or irrationality of justifications without checking motivating reasons.

Consider, finally, a common example that Audi mentions (1985, 159–160): “a person cites an altruistic reason he had for helping someone, when in fact he was motivated by selfish reasons.” If what explains that action is selfish reasons, one would expect that the person would not behave the

² I say that there *might* be an incoherence because I am not sure that there is no relevant difference between the two scenarios to which the agent could rightly point out. After all, if there are many onlookers, the agent could always argue that she thought that someone would take care of the person in distress, and perhaps that could be a legitimate expectation.

same way in a situation in which she again must help someone but the selfish reasons are absent—there is no benefit for her. That would reveal an incoherence in her attitude towards purported normative reasons between both cases. Otherwise, if her behaviour is *consistently* helpful, it seems to me that insisting on the existence of a selfish motivation in order to criticise her reasons would entail a moral theory that is way too demanding.

All this is not intended to mean that we should *never* take into account motivating reasons when assessing justifications. Even within the boundaries of a single action, *sometimes* a manifest mismatch between purported normative and motivating reasons can be reprehensible. If it is clearly apparent, for example, that I intended to punch someone out of anger and, by sheer luck, I ended up moving him away from a bus that was going to run over him, thus saving his life, then I can hardly justify my action by saying that I saved his life. Anyone could see that my intention was to hit him. However, apart from clear cases like this one, our practice of giving and asking for justifications should not focus on mismatches between purported normative, motivating and explanatory reasons. We should accept that those mismatches are ubiquitous in human action, as the research in experimental psychology has shown, but at the same time we can be confident that we have the resources to assess justifications by themselves.

5. Conclusion

Research in cognitive psychology during the last five decades has shown that, in many situations, the reasons with which people explain their own actions and judgements do not correspond to the real factors that caused them. This finding has led to the conclusion that people do not have introspective access to the causes of their own behaviour; instead, people infer them, just as they would if they were observers of someone else's behaviour. Such a conclusion seems to cast doubt on the significance of our practice of justification of actions and exchange of reasons. However, in order to fully understand the philosophical implications of the results of psychological research, we need to be clear about what kinds of reasons psychologists are talking about.

In the philosophical literature, three kinds of reasons have been distinguished, according to the kind of question that they answer: *normative* reasons, considerations that make an action right; *motivating* reasons, considerations in light of which the person acted; and *explanatory* reasons, considerations that explain what caused an action or an event. The problem with the psychological experiments, we saw, was that the participants offered purported motivating reasons that did not explain their choices at all; instead, what explained their choices were explanatory reasons that the experiments uncovered and of which the participants were unaware. That challenges the reality of motivating reasons, and we are left only with normative reasons that, for all we know, could have no effect on behaviour whatsoever—they could be epiphenomenal.

However, we also saw that certain behaviours could only be plausibly accounted for by the influence of purported normative reasons. If our performance in a logical task is better when there is argumentation, and if we tend to lean towards the most justifiable decisions when our intuitions about what to do are weak, that gives us grounds for believing that sometimes purported normative reasons do guide our actions. If that is the case, then actual normative reasons—being a subset of purported normative reasons—must at least sometimes influence our behaviour. The problem, given our lack of introspective access to the causes of our behaviour, is that in practice we can never be sure that, in a particular instance, we are genuinely motivated by normative reasons. For this reason, I argued that our assessments of the reasons produced by agents should not give much weight to whether they are also motivating reasons or not—i.e. whether they are rationalisations of actions. Outside laboratory conditions, the identification of motivating reasons is a tricky issue and it is bound to lead to speculations, and we have the conceptual resources to assess purported normative reasons in themselves.

Acknowledgements

A previous version of this paper was presented at the 12th Conference of the Ontario Society for the Study of Argumentation (OSSA), which took place on-line on 3-6 of June of 2020. Special thanks to my commentator, Marcin Koszowy, for his useful observations. I should also thank the people who

attended my talk and offered insightful suggestions, especially Daniel Cohen, Hans Hansen and Júlder Gómez.

Funding

This research was possible thanks to the postdoctoral scholarship FOND-ECYT 3190149 of ANID/CONICYT, and also to the project PGC2018-095941B-I00, “Prácticas argumentativas y pragmática de las razones,” of the Spanish Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación.

References

- Alvarez, Maria. 2009. “How Many Kinds of Reasons?” *Philosophical Explorations* 12 (2): 181–93. <https://doi.org/10.1080/13869790902838514>
- Audi, Robert. 1985. “Rationalization and Rationality.” *Synthese* 65 (2): 159–84. <https://doi.org/10.1007/BF00869298>
- Cohen, Daniel H. 2019. “Argumentative Virtues as Conduits for Reason’s Causal Efficacy: Why the Practice of Giving Reasons Requires That We Practice Hearing Reasons.” *Topoi* 38 (4): 711–18. <https://doi.org/10.1007/s11245-015-9364-x>
- Dancy, Jonathan. 2000. *Practical Reality*. New York: Oxford University Press.
- Davidson, Donald. 1963. “Actions, Reasons, and Causes.” *Journal of Philosophy* 60 (23): 685–700. <https://doi.org/10.2307/2023177>
- Evans, Jonathan St. B. T. 2010. *Thinking Twice: Two Minds in One Brain*. New York: Oxford University Press.
- Frederick, Shane. 2005. “Cognitive Reflection and Decision Making.” *Journal of Economic Perspectives* 19 (4): 25–42. <https://doi.org/10.1257/089533005775196732>
- Haidt, Jonathan. 2001. “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment.” *Psychological Review* 108 (4): 814–34. <https://doi.org/10.1037/0033-295x.108.4.814>
- Kiesewetter, Benjamin. 2017. *The Normativity of Rationality*. Oxford: Oxford University Press.
- Kjeldsen, Jens E. 2020. “What Makes Us Change Our Minds in Our Everyday Life? Working through Evidence and Persuasion, Events and Experiences.” In *OSSA Conference Archive*, 1–14. <https://scholar.uwindsor.ca/ossaarchive/OSSA12/Saturday/7>.

- Kunda, Ziva. 1999. *Social Cognition: Making Sense of People*. Cambridge, MA: MIT Press.
- Latané, Bibb, and John M. Darley. 1970. *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century Crofts.
- Mercier, Hugo, and Dan Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press. <https://doi.org/10.4159/9780674977860>
- Nisbett, Richard E., and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, NJ: Prentice-Hall.
- Nisbett, Richard E., and Timothy DeCamp Wilson. 1977. "Telling More than We Can Know: Verbal Reports on Mental Processes." *Psychological Review* 84 (3): 231–59. <https://doi.org/10.1037/0033-295X.84.3.231>
- Parfit, Derek. 1997. "Reasons and Motivation." *Proceedings of the Aristotelian Society, Supplementary Volumes* 71: 99–130. <https://doi.org/10.1111/1467-8349.00021>
- Scanlon, Thomas M. 2014. *Being Realistic about Reasons*. New York: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199678488.001.0001>
- Searle, John R. 2001. *Rationality in Action*. Cambridge, MA: MIT Press.
- Siegel, Harvey. 1997. *Rationality Redeemed? Further Dialogues on an Educational Ideal*. New York: Routledge.
- Trouche, Emmanuel, Emmanuel Sander, and Hugo Mercier. 2014. "Arguments, more than Confidence, Explain the Good Performance of Reasoning Groups." *Journal of Experimental Psychology: General* 143 (5): 1958–71. <https://doi.org/10.1037/a0037099>
- Wegner, Daniel. 2002. *The Illusion of Conscious Will*. Cambridge, MA: MIT Press. <https://doi.org/10.7551/mitpress/3650.001.0001>
- Wilson, Timothy D. 2002. *Strangers to Ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard University Press.
- Wilson, Timothy de Camp, and Richard E. Nisbett. 1978. "The Accuracy of Verbal Reports about the Effects of Stimuli on Evaluations and Behavior." *Social Psychology* 41 (2): 118–31. <https://doi.org/10.2307/3033572>

A Dilemma about the Mental


Guy Dove* – Andreas Elpidorou**

Received: 24 April 2019 / Revised: 8 June 2020 / Accepted: 11 September 2020

Abstract: Physicalism demands an explication of what it means for something to be physical. But the most popular way of providing one—viz., characterizing the physical in terms of the postulates of a scientifically derived physical theory—is met with serious trouble. Proponents of physicalism can either appeal to current physical theory or to some future physical theory (preferably an ideal and complete one). Neither option is promising: currentism almost assuredly renders physicalism false and futurism appears to render it indeterminate or trivial. The purpose of this essay is to argue that attempts to characterize the mental encounter a similar dilemma: currentism with respect to the mental is likely to be inadequate or contain falsehoods and futurism leaves too many significant questions about the nature of mentality unanswered. This new dilemma, we show, threatens both sides of the current debate surrounding the metaphysical status of the mind.

* University of Louisville


 <https://orcid.org/0000-0003-0470-7006>

 Department of Philosophy, University of Louisville, 313 Bingham Humanities Building, Louisville, Kentucky 40292 USA

 guy.dove@louisville.edu

** University of Louisville

 <https://orcid.org/0000-0001-7317-339X>

 Department of Philosophy, University of Louisville 313 Bingham Humanities Building, Louisville, Kentucky 40292 USA

 andreas.elpidorou@louisville.edu



Keywords: Consciousness; materialism; mental; metaphysics; mind; physical; physicalism; reduction; theory.

1. Introduction

Physicalism faces what is known as “Hempel’s Dilemma.”¹ This dilemma emerges in an attempt to answer the question of how we are to characterize the physical. Not to put too fine a point on it, we can either choose to characterize it in terms of some current theory or some future theory. Those who support the second option typically appeal to a complete and ideal future theory. Neither option seems particularly promising: Choosing current physical theory would almost assuredly make physicalism false or incoherent,² and choosing a future theory would seem to render physicalism indeterminate or perhaps trivial.³

The purpose of this essay is to argue that a similar dilemma threatens any (non-eliminativist) approach to the mental that attempts to answer significant metaphysical questions about the status of the mind. The idea itself is fairly straightforward: Currentism with respect to the mental is likely to lead to false claims and futurism leaves too many questions unanswered. Insofar as a metaphysical position takes the content of “mental” as either settled or unproblematic, it will falter against the Scylla and Charybdis of this dilemma. The dilemma thus threatens the foundations of the current debate surrounding the metaphysical status of the mind.

¹ Hempel (1969; 1980), but see also Chomsky (2000), Crane and Mellor (1990), and Melnyk (1997).

² If “physical” means the posits of current physics, then physicalism—the view that holds that everything that exists is physical—is false because the inventory of current physics is incomplete. In addition, if understood in terms of current physics, physicalism is likely incoherent because of the existing inconsistencies between the subfields of physics (Wilson 2006)

³ We do not know what such a future theory would look like, nor do we know whether it will end up positing mental entities as fundamental. The fact that it is not possible to determine which of these options—indeterminacy or triviality—will obtain just further highlights the epistemic challenges facing those who embrace the second horn of the dilemma.

We want to be absolutely clear about what is, and is not, on offer. The dilemma provides neither a positive argument for physicalism nor a refutation of anti-physicalism. The challenge posed by the existence of phenomenal consciousness to physicalist attempts to understand our place in the world does not magically disappear by acknowledging the difficulties that characterizations of the mental face. What the dilemma does is to expose an important blind spot shared by many standard positions in the philosophical discussion of the metaphysical status of the mental. Many participants in this discussion acknowledge the need for precision and nuance when it comes to articulating the ways in which the mental could be related to the physical. It is precisely due to such a concerted and sustained effort to better understand the relationship between the mental and the physical that the literature has been populated by attempts to describe and refine the notions of *identity*, *reduction*, *supervenience*, *realization*, *emergence*, and *grounding* and apply them to the mind-body problem.⁴ Furthermore, both proponents and critics of physicalism have rightly paid much attention to the nature of the physical, asking what it is and how it can be defined.⁵ Unfortunately, however, the very same participants often fail to apply the same kind of rigor and questioning attitude when it comes to the nature of the mental as it figures in the mind-body (or brain) problem. Instead, they typically appeal to intuitive, rough-and-ready characterizations of the mental (*qua* mental phenomenon) that are thought to suffice for the purposes of examining the metaphysical status of mentality.

There are *prima facie* reasons to question the adequacy of such characterizations. Consider the case of vitalism. Historically, the phenomenon of life was thought to provide a clear exception to materialism. One factor contributing to vitalism's demise was the inability of its supporters to settle on a precise characterization of vital forces (Mayr 1982). The comparison between consciousness and life has on occasion been dismissed as little more

⁴ The literature is too expansive to review here. For recent surveys, see Elpidorou (2017), Stoljar (2015), and Tiehen (2018).

⁵ See, e.g., Bokulich (2011); Dove (2016); Dowell (2006); Melnyk (1997) and (2003); Montero (2001) and (2009); Montero and Papineau (2005); Ney (2008); Spurrett and Papineau (1999); Stoljar (2001); Tiehen (2016); Vicente (2011); Wilson (2006); Witmer (2016); and Worley (2006).

than a weak argument from analogy (e.g., Chalmers 2003). However, the overall success of physicalist explanations of phenomena that were previously thought to be exceptions to physicalism (Melnyk 2003) and the apparent causal closure of the physical (Papineau 2001) suggest that a more robust account of the mental might be required, just as a richer account of life was required. To put the same point somewhat differently, we need to have reasons that are not shaped by retrospective bias to think that our ideas about the nature of the mental are more solid than our past ideas about life. In the absence of such reasons, the rich, robust, and diverse circumstantial case for physicalism weighs heavily and forces us to take seriously the idea that there might be more to the mental than what our intuitive characterizations of it reveal.

Our aim in this paper is to demonstrate why an appeal to an intuitive, pre-theoretical notion of the mental is insufficient for the purposes of investigating the ontological status of the mental and specifically, of consciousness. Although a comparison between vitalism and physicalism (or between the concept of *life* and the concept of *mental*) is suggestive, our case does not rest on that comparison. In fact, we will show that there are reasons internal to the debate pertaining to the ontological status of consciousness that support the need for a rich account of the mental. Because of that, characterizations of the mental—just like those of the physical—face a choice between currentism or futurism.

There has been remarkable progress in both scientific and philosophical investigations of the nature of mentality. Precisely because of such progress, the nature of mentality cannot be assumed to be an unproblematic given. Indeed, we have not reached a settled understanding of the mental yet, nor do we know enough to predict confidently how future theories will describe the mental. But if the meaning of “mental” as this figures in the mind-body problem is unclear, not fully understood, or subject to revision and change, as we will argue that it is, then claims about the relationship between the mental to the physical would also be unclear, not fully understood, or subject to revision and change. According to our view, progress in resolving the mind-body problem is unlikely to take place without first acknowledging that we have only a limited grasp of the nature of mentality.

2. Why is this a New Dilemma?

Given that our dilemma mirrors Hempel's Dilemma to some degree, one might wonder why there has been almost no discussion of it to date.⁶ One reason may be the perception that anti-physicalism and non-physicalist views just do not face the same sort of questions as physicalism. Giving voice to this perception, Levine and Trogon (2009, 356) write:

A longstanding issue in the philosophy of mind is how to specify the sense of "physical" at issue with materialism. There is no corresponding problem, however, for specifying mentality; mental properties are either conscious properties or intentional properties.

We do not take Levine and Trogon's remark as proof that there is an agreed-upon characterization of the mental. What the quoted passage suggests instead is that, compared to the task of articulating the physical, describing the mental is an easier task. In the case of the mental, we have some grasp of the essence of mentality: whatever the mental is, it is either the phenomenal or the intentional, or both. Such an understanding of the mental can then serve as our starting point in sketching out the various positions in the debate surrounding the ontological status of the mind and its relationship to the brain, body, and world.

As a matter of actual philosophical practice, there appears to be little disagreement concerning how to broadly define the mental. Although we do not quarrel with the cultural or sociological significance of this claim, we do challenge the notion that such general agreement regarding the mental settles anything. Consider, for example, what would happen if all the philosophers who support physicalism got together and agreed once and for all that the physical should be defined in terms of a commitment to Cartesian corpuscles. Ex hypothesis there would be no disagreement concerning how to characterize the physical. But this would not in any way remove the

⁶ The only explicit discussion of this new dilemma that we have found is in (Gillett and Witmer 2001). Tellingly, they dismiss it straight away. They contend that our special epistemic access to mental entities blocks the dilemma. See section 3 below for a discussion of this approach.

challenge posed by Hempel's Dilemma, which has nothing to do with philosophical agreement and everything to do with accuracy and truth.

Reflecting a kind of philosophical common sense (at least within analytic philosophy), Levine and Trogon offer a disjunctive characterization of the mental, defining it in terms of either conscious or intentional properties. Because conscious properties are central to so many of the important debates concerning the mental and the physical, we are going to focus on them exclusively in this essay. However, the same arguments that we employ with respect to characterizing the mental in terms of conscious properties can be employed *mutatis mutandis* to attempts to define it in terms of intentionality.⁷ When we turn to questions concerning the nature of conscious properties, we find that there is a great deal of disagreement concerning how to define them. In other words, all that is accomplished by the disjunctive characterization offered above is that it pushes the problem down a level. Or so we will argue in section 3.

An additional reason for the lack of consideration of our proposed dilemma is that claims about the mental play a different role in the debate over physicalism than our claims about the physical. At a minimum, physicalism rests on the universally quantified claim that all relevant phenomena, including those that we identify as mental, are ultimately physical. Under the typical rough-and-ready formulation, physicalism holds that there is nothing over and above the physical. Hempel's concerns our purported inability to arrive at a characterization of the physical that is able to support this universal claim. In particular, supporters of the dilemma focus on our inability to rule out fundamental mental properties, entities, events, etc. that would seem to violate physicalism. Given this, a strictly analogous dilemma would apply to idealism or what might be called mentalism (the claim that there is nothing over and above the mental). Of course, contemporary adherents of this sort of metaphysical position are hard to find. Instead, anti-physicalists tend to defend the existentially quantified claim that particular mental properties, entities, events, etc. are ultimately not physical ones. Thus, if our dilemma is to be relevant to contemporary debates, it needs to undermine our confidence in our ability to

⁷ Indeed, we suggest that it will likely be easier to make the case with respect to intentional properties because their theoretical nature is more apparent.

characterize mental phenomena in such a way that confirms or disconfirms this more circumscribed metaphysical thesis. In keeping with this, we endeavor to show that our dilemma does indeed undermine confidence in such claims. We argue that neither our current understanding of the mental nor our projectible future understanding is up to the job of settling such metaphysical questions.

The lack of discussion of our proposed dilemma could also be due to the fact that it is a late entry into a crowded field. We already know that the mental presents special philosophical challenges. After all, there is an extensive literature on a number of well-established problems concerning the mental—including the relationship, causal or otherwise, between the mind and body (Jackson 1982; Chalmers 1996; Kim 1998; Libet 1985; Robb and Heil 2013), the hard problem of consciousness (Chalmers 1996), the knowledge argument (Jackson 1982, 1986), and the explanatory gap (Levine 1983; Nagel 1974). Do we really need a new problem concerning the mental? We think that we do. One of the reasons that we think this is that the problem identified in this essay is very different than the other better-known problems. Indeed, it creates conceptual challenges for most of the currently identified problems because they generally view the mental as a serious threat to physicalism or, more broadly, naturalism. Following Jackson (1998), we can categorize most of the traditional puzzles as “location problems” where the core issues are concerned with how to place mental entities or properties in a physical or natural world. These problems generally depend on accepting certain positive claims about the nature of the mental. The new dilemma—which we will heretofore refer to as “DaM” for Dilemma about the Mental—seems to make both the formulation and solution of location problems harder.

Lastly, one might think that DaM does not apply to the mental because of the general perception that the sources of Hempel’s original dilemma are the very features of physicalism that supporters of alternative metaphysical accounts of the mental oppose, such as commitments to metaphysical naturalism, reductionism about the mental,⁸ and perhaps some variety of

⁸ Here we are treating reductionism as the position that holds that the mental is nothing over and above the physical.

scientism.⁹ Because many metaphysical stances on the mental involve an explicit disavowal of these features,¹⁰ it is not surprising that their supporters would not worry about an analogous dilemma.¹¹ We suggest that this perception is mistaken. Hempel’s Dilemma primarily arises from two related elements. The first is a set of reasons to think that our current understanding of the physical is incomplete or inadequate. Certainly the fact that, taken as a whole, contemporary physics is inconsistent counts as a red flag (Wilson 2006). The second is a set of reasons to think that our future understanding will be theoretically transformative in ways that are difficult to appreciate from our current epistemic standpoint. A history consisting of profound theoretical upheavals with respect to our conception of the physical provides an inductive case for the likelihood of radical future theoretical innovation.

We propose that analogs of these two elements—neither of which requires a commitment to naturalism, reductionism, or scientism—are present with respect to the mental. We take our current understanding of the mental to be, at the very least, significantly incomplete or inadequate. And we believe that a future understanding of the mental is likely to be theoretically transformative. Consider, for example, what the possibilities of conscious AI, teleportation, or brain-to-brain communication could teach us about the mind.

3. Troubles with currentism

In this section, we set out to accomplish two things. First, we argue that there is actually substantial disagreement concerning the nature of

⁹ The term *scienticism* is often used with a negative connotation. However, there has been a recent effort to reclaim the positive sense of this term (e.g., Ladyman and Ross 2007) in much the same way that some philosophers of mind have sought to reclaim the positive sense of *reductionism* (e.g., Churchland and Churchland 1992).

¹⁰ It is important to recognize that this characterization is not universal. For example, Chalmers (1996, 128) defends a form of dualism that “...is naturalistic because it posits that everything is a consequence of a network of basic properties and laws.”

¹¹ Physicalists, on the contrary, are not particularly focused on finding new problems concerning the mental.

conscious properties. Second, we contend that this disagreement throws currentism into question.

3.1. Historical precursors

Before we get to our argument, we want to acknowledge that our focus on the question of how to characterize conscious properties is not without precedent. We are not the first to highlight the philosophical importance of this issue. Consider a well-known anecdote concerning a public interaction between the philosophers Herbert Feigl and Rudolph Carnap (we are relying on Tim Crane's 2007 recounting of this event). In the course of a talk defending physicalism at UCLA, Feigl admitted that science had yet to provide a physical explanation of the qualia associated with phenomenal experience. Carnap, who was in the audience, purportedly interrupted Feigl, and the following exchange is supposed to have happened (Crane 2007, 16-17):

- Carnap: But Feigl, there is something missing from your lecture. Science is beginning to explain qualia in terms of the alpha factor!
- Feigl: Carnap, please tell me: What is the alpha factor?
- Carnap: Well, Feigl if you tell me what qualia are, I'll tell you what the alpha factor is.

Whether or not this conversation actually occurred in this manner, we agree with Crane that the point about qualia is well taken.¹² Too many philosophical discussions about consciousness in general and phenomenal conscious properties in particular rest on the implicit assumption that these phenomena are well understood.

Carnap's rejoinder fits with the general positivist emphasis on the need for philosophers to clearly define their terms, preferably in a way that meshes with the physical sciences. We raise the question of how to characterize conscious properties for different reasons than Carnap: we are not positivists, we are not defending physicalism, and we are not dismissing the

¹² In another article, Crane (2001, 170) laments: "To have a clear understanding of this problem, we have to have a clear understanding of the notion of qualia. But despite the centrality, it seems to me that there is not a clear consensus about how the term 'qualia' should be understood, and to this extent the contemporary problem of consciousness is not well-posed."

relevance of conscious properties. We don't even fully agree with Carnap about the facts on the ground. In contrast to his assessment, we think that philosophers have in fact tried to tell us what conscious properties are. Indeed, a lot of work has been done on this topic in the intervening decades. The trouble is that these philosophers often provide very different answers, and their disagreement threatens the first horn of the dilemma.

3.2. The center will not hold

To date, we have found only one explicit discussion of something like DaM: Gillett and Witmer (2001) acknowledge the prima facie plausibility of such a dilemma but then argue that it is blocked by our special epistemic access to mental entities. However, special epistemic access is not enough to block DaM because even if we grant that such access exists this would not guarantee that one possesses an adequate understanding of the mental. This seems particularly true of consciousness itself, which of course lies at the center of the disagreement over the status of mental entities. Chalmers (1996, 3) himself notes that consciousness, “can be frustratingly diaphanous: in talking about conscious experience, it is notoriously difficult to pin down the subject matter.” Indeed, the claim that we have some kind of special or privileged access to our mental states often reduces to one or both of the following claims: (i) Mental states are self-luminous (if a subject is in mental state M, then the subject knows that they are in M) and (ii) Mental states are incorrigible (if a subject believes that they are in mental state M, then they are in M). Neither claim suffices to show that we know the nature of our mental states.

Even if one restricts the mental to (phenomenally) conscious properties, our understanding depends upon our conceptualization of such properties. What this shows is that what we take to be features of conscious properties depends on the nature of our phenomenal concepts (i.e., the concepts that we use when we introspectively examine the phenomenal character of our experience). If the deployment of phenomenal concepts in introspection reveals to us the entire¹³ nature of their referents, then currentism is safe. But

¹³ If phenomenal concepts reveal only part of the essence of their referents, then we could still be mistaken about the nature of consciousness and qualia.

why should one accept such a strong claim? The use of phenomenal concepts may reveal only *part* of the essence of their referents, or it may only reveal accidental features of their referents (ones that allow us to uniquely identify them in the actual world), or it may fail to reveal *any* features at all.¹⁴ If any of these three possibilities could be true, then it would undermine currentism.

Moreover, there are very good reasons to be concerned about our current understanding of conscious properties. Their precise nature has been—to echo the quote from Chalmers given above—notoriously difficult to pin down. Responding to the question of what qualia are, Block (1980, 278) famously quipped “As Louis Armstrong said when asked what Jazz is, ‘if you got to ask, you ain’t never gonna get to know’.” The trouble is that a number of people have asked and, more importantly, given different answers.¹⁵ Some categorically deny the existence of qualia at all (Dennett 1991). Others deny that qualia exist where one might think that they should. For instance, Tye and Harman, utilizing the supposition that experiences are transparent or diaphanous, have claimed that qualia are not properties of our visual experiences (e.g., Harman 1990; Tye 2000; cf. Dretske 1995).¹⁶ Some make the controversial claim that qualia are (or reduce to) the representational contents of our experiences (e.g., Dretske 1995; Lycan 1996). Others hold instead that qualia are intrinsic, non-representational properties of our experiences (e.g. Block 1990; Peacocke 1983). Yet others hold a relational (i.e., direct realist) account of qualia (e.g., Campbell

¹⁴ See Nida-Rümelin (2006) and Goff (2011) and critical discussions in Diaz-Leon (2014), Elpidorou (2016) and Trogdon (2016).

¹⁵ Our discussion of qualia in this section follows that of Crane (2001). We recommend looking to his paper for a richer and more detailed discussion of the diversity of opinion that exists within the philosophical literature concerning this fundamental notion.

¹⁶ Tye (2017) explains: “[Q]ualia, conceived of as the immediately ‘felt’ qualities of *experiences* of which we are cognizant when we attend to them introspectively, do not really exist. The qualities of which we are aware are not qualities of experiences at all, but rather qualities that, if they are qualities of anything, are qualities of things in the world (as in the case of perceptual experiences) or of regions of our bodies (as in the case of bodily sensations). This is not to say that experiences do not have qualia. The point is that qualia are not qualities of experiences.”

2002; Brewer 2011). Finally, there is even disagreement as to what organisms undergo states with qualia. Do insects, for instance, experience qualia?

The aforementioned disagreements are substantial. If we do not know where on the phylogenetic scale qualia start to appear, then we do not have a good understanding of the necessary biological conditions for their existence. If we do not know whether qualia should be individuated narrowly or widely, then we do not know what kind of contribution (causal or constitutional) the world makes. And, most importantly, if we cannot agree whether qualia are properties of experiences or not, then it is unclear whether we have really understood the notion of *qualia*.

But isn't there something intuitively clear and distinctive about our conscious experiences (e.g., Chalmers 2010)? Don't we know that in some sense or another qualia must exist? Don't we know what the redness of the setting sun is? If so, isn't *that*, admittedly minimal, and pre-theoretical understanding of qualia sufficient to furnish us with a satisfactory account of mentality? The answer, we believe, is simply "No." As noted by both Crane (2001) and Keely (2009), many of the same claims that are made in support of qualia were made about the currently disfavored notion of *sense-data*. For example, Price in an article from 1932 (quoted by Crane 2001, 175) explains:

The term sense-datum is meant to be a *neutral term* ... The term is meant to stand for something whose existence is indubitable (however fleeting) something from which all theories of perception ought to start.

Now the mere fact that many of the same, controversial claims made about *qualia* were made about the earlier notion of *sense-data* does not in and of itself show that the former are false, but it does suggest that more is needed to settle the matter than a careful self-examination of our inner experience.

Furthermore, the minimal understanding of qualia that one is able to find when one introspects on one's experiences is insufficient to answer the many questions that would allow us to expose the nature of qualia. Can one, simply by focusing on the painfulness of pain or the redness of a Rothko painting, settle whether qualia are properties of experiences or not? Can they determine whether the environment or our bodies make a constitutive contribution to the content of our experience? Can they tell whether it is a possible for a physical and functional duplicate of an acrophobic subject to

fail to experience fear when they stand near the edge of a tall building? Can they resolve the issue of whether qualia are physical or not? By itself, a minimal, pre-theoretical understanding of qualia, if such a thing exists, cannot settle the nature of mentality. This is evident not only by the vast array of diverging and in some cases contradicting accounts of qualia, all of which seem to appeal to this pre-theoretical notion of qualia (see, e.g., Tye (2017) for an overview), but also by the fact that attempts to specify further the nature of qualia require substantial assumptions about the nature of introspection, awareness, representation, and concepts.

One could argue that our focus on qualia is somewhat dated. In other words, one could possibly claim that our discussion of currentism is not current enough. Current discussions of conscious properties—so the argument might go—are not as reliant on either the term or the concept of qualia. We are willing to grant the possibility that such a shift might be underway. One might even be able to find quantitative evidence for the waning influence of the concept of qualia (tied perhaps to the decreased use of this term). This, however, would not be enough to undermine our general point. What needs to be shown to do this is that significant theoretical disagreement about the precise nature of conscious properties does not exist. We hold the line here and maintain that, if anything, the level of disagreement has increased. In fact, it seems likely that the very reason for the emerging distaste for the term *qualia* (a distaste which is by no means universal) is a lack of agreement about the nature of qualia. Moreover, it is hard not to see how this purported shift away from qualia is not grist for our mill. After all, if the notion of qualia is currently falling out of favor the way that sense-data fell out of favor decades ago, then we have some inductive reason to question our current understanding. Again, this is not an argument that we will never arrive at a successful theory. More importantly, it is not an attack on the effort to theorize about the nature of conscious properties. Instead, it is merely pointing out there is reason to think that our current understanding is not accurate or secure enough to settle important metaphysical questions.

We propose that the existence of significant disagreement about the very nature of conscious properties impugns our current state of knowledge. This is not a denial of the existence of conscious properties but rather an

assessment of our current understanding of them. Despite appearances to the contrary, our current understanding of mentality (at least as it is defined relative to conscious properties) is far from complete and likely to be mistaken in significant ways. Future developments in philosophy, biology, neuroscience, and psychology could render many of our beliefs about mentality false.¹⁷

4. The prospects of futurism

Above we argued that our current understanding of mentality leaves much to be desired. Could one characterize the mental in terms of a final and complete theory or account of the mental? Just as any meaningful form of physicalism must avoid positing *sui generis* non-physical entities, any meaningful and positive account of the character and ontology of the mental must avoid a robust eliminativism in which the mental does not exist. We believe that a futurist approach is fraught with difficulties. Our main contention is that, given our current epistemic standing in regard to mentality, the shape of this final theory or account is severely and problematically (for present purposes) indeterminate. This is due primarily to two facts. First, there are a number of competing theories of the mental that might turn out to be true. Second, most (if not all) of these possible final theories are radically transformative insofar as their success would require a substantial revision of our current understanding of the mental. The question remains: Does this revision amount to a wholesale rejection of the mental as it is currently understood or not? We suggest that we do not know enough about the future outcomes of our theories of mentality to be in a position to make an informed judgment about their content.

Consider for instance the following possible final theories about the mental: emergentism, non-reductive physicalism (e.g., realization), neutral monism, property dualism, panpsychism, or something completely new and

¹⁷ Scientific discoveries, e.g., blindsight (Weiskrantz 1980) and the two visual streams hypothesis (Milner and Goodale 2006), and conceptual advances, e.g., the phenomenal/access consciousness distinction (Block 1995) have transformed our understanding of consciousness and thus of mentality.

unknown. Obviously the last option is a non-starter as a means of avoiding indeterminacy. But all the other ones are also problematic. Indeed, many of the problems that beset a *currentist* attempt to define the mental also beset these future theories. As we saw in the previous section, our current understanding of conscious properties is at best incomplete. On the one hand, there is significant disagreement about the nature of conscious properties, and on the other hand, any understanding of conscious properties that can serve as the “common denominator” across different positions appears to be incapable of settling substantial questions about the character of the mental. Consequently, any attempt to characterize the mental on the basis of a future theory that utilizes our *current* understanding of qualia would be problematic insofar as the content of such a future theory is (given our current epistemic perspective) severely indeterminate. We just do not know which of the many competing accounts of the mental we ought to accept, and if we opt for a minimal (or “thin”) understanding of the mental, then our future theory will fail to specify the ontological status of the mental. The point is simple: we don’t know enough about qualia right now to be able to draw meaningful conclusions concerning how the notion of qualia will be developed in the future.

Finally, it will not work to characterize the mental in terms of a final theory but not specify whether that theory is, e.g., panpsychism, neutral monism, or realization physicalism. Each of them tells us something radically different about the nature of mentality. Put crudely, the first holds that the mental is fundamental, the second holds that fundamentally nothing is mental, and the third holds that the mental exists but only derivatively (it is nothing over and above the physical). The fact that some of our candidates for a final theory of the mental are deeply at odds with each other shows that before we can define the mental by using one of these theories, we have to decide which one is *likely* to be true. But if philosophical debates about the nature of consciousness are any indication, we are far away from being able to do so. All in all, the prospects of futurism appear to be rather dim.

5. Objections and replies

We have argued that attempts to characterize the mental encounter a dilemma similar to the one faced by physicalists in their efforts to define the physical. Given our contemporary understanding of the mind—including all that we have gathered through our phenomenological experiences, philosophical investigations, and the psychological and brain sciences—neither currentism nor futurism with respect to consciousness and other purportedly non-physical mental properties holds much promise. Recognizing that DaM flies in the face of philosophical conventional wisdom and threatens to upend established philosophical debates concerning the relationship of the mental to the physical, we review some possible objections and offer responses below. The objections share a common theme: the idea that the dilemma, for one reason or another, does not apply to the debate between physicalists and anti-physicalists.

Before getting to these objections though, we want to emphasize what we see as the fundamental force of the dilemma. We suggest that the dilemma throws into question our capacity to answer important metaphysical questions surrounding consciousness and other mental phenomena given our current knowledge. It thus decidedly does not amount to a defense of either physicalism or anti-physicalism. What it does do is threaten the arguments offered in support of either metaphysical position. In other words, we acknowledge that identifying the dilemma does not make consciousness any less mysterious or help us to see how it fits in our world. Indeed, the dilemma demonstrably adds to the mystery. After all, it demonstrates that we know even less we thought we did about conscious properties. DaM in no way forces us to deny the presence of the significant epistemic gap that inspires most of the important philosophical puzzles. Having said this, we need to also point out that DaM becomes an issue for *specific* formulations of these puzzles. If it obtains, then our epistemic grip on the mental may well be insufficient to affirm the possibility of inverted spectra or philosophical zombies, to arrive at an adequate understanding of the super-scientist Mary, or to outline clearly the hard problem of consciousness. In other words, DaM places pressure on so-called location problems by throwing into question our understanding of what is being located.

5.1. *Objection #1*

There is an important asymmetry between the roles that the terms “physical” and “mental” (or “qualia”) play in our evaluation of physicalist theories of mind. “Mental” refers to the properties, entities, events, etc. that need to be explained whereas “physical” denotes the set of properties, entities, events, etc. that are meant to be doing the explaining. Because of the role of “physical” we naturally need a precise account of it. However, because of the non-explanatory role of “mental” we do not need nearly as much precision as we need with “physical.” All that we need is some rough understanding of mentality—one that perhaps can be given ostensibly: mentality is THAT!

Reply: We suggest that this objection turns on a misunderstanding of the force of the dilemma. To see why, consider the role that Hempel’s original dilemma plays in the debate. Few suggest that, because of the dilemma, we should eliminate our everyday notion of physical properties, entities, events, etc. More to the point, few suggest that we should stop doing physics because of the dilemma. What people do suggest is that the dilemma shows that we do not know enough about the ultimate nature of the physical to answer important metaphysical questions—in particular, the question of whether or not mental entities or properties are physical. In an analogous fashion, this new dilemma should not be seen as an attack on our everyday notion of the mental or indeed on our everyday conception of conscious properties.¹⁸ Nor should it be seen as attempt to preempt philosophical or scientific investigation of these. As was the case with the original dilemma, DaM threatens attempts to settle the relevant metaphysical questions by an appeal to a substantial account of mentality.

Returning to the objection, we can make the issues raised by DaM explicit. We have no problem granting that there are everyday conceptions of phenomenal experience or other mental phenomena that serve as a starting point for philosophical investigation. These phenomena may well serve in some sense as explananda for philosophical explanation (although we would note that in most non-dogmatic areas of human investigation it is not

¹⁸ Of course, suggesting that there is an everyday notion of qualia is contentious to say the least. This fact alone seems to offer support to our main position.

uncommon to modify one's understanding of what is to be explained in the course of developing explanations).¹⁹ We do not even question that they may raise intriguing metaphysical location problems. We do suggest, though, that accounts of such phenomena that are rich enough to address the relevant metaphysical questions will need to offer a meaningful characterization of their mentality. As Robert van Gulick notes in his discussion of the explanatory gap, "the more we can articulate the structure within the phenomenal realm, the greater the chances for physical explanation; without structure we have no place to attach our explanatory 'hooks'" (1997, 565). Thus, attempts to determine whether an aspect of our conscious life can be explained in physical terms must begin with detailed descriptions of that aspect. It is here that the dilemma becomes relevant, for as we argued in section 3, our understanding of the mind appears to be both incomplete and likely to be mistaken in significant ways. Importantly, this problem arises independently of any understanding of the physical. For instance, it would arise within the context of a full-throated idealism where everything that exists is mental.

5.2. *Objection #2*

Because anti-physicalism is defined in terms of its opposition to physicalism, it is not undermined by DaM.

Reply: We freely admit that DaM arises in the context of positive claims about the nature of mentality. This raises the possibility that an anti-physicalism devoid of such claims could elude its reach. We suggest however that it is very difficult to envision a substantial form of anti-physicalism that is free of positive commitments concerning the character of the mental. Given that many arguments for anti-physicalism depend crucially on observations concerning conscious experience and a number of philosophers of science have questioned the very notion of theory-neutral observation

¹⁹ In the case of consciousness, P.S. Churchland (1986), P. M. Churchland (1995), and Flanagan (1992), among others, have argued for a co-evolutionary approach to the problem of consciousness: one that simultaneously examines the problem of consciousness both from the physical (biological, neuronal, or bodily) and the mental (or phenomenal) perspective.

(Azzouni 2004; Bogen 2016; Chang 2005; Duhem 1906; Hanson 1958; Kuhn 1962), it seems reasonable to suppose that theoretical influences—perhaps implicit ones—may be at work. More importantly, as became apparent in our discussion of the current debates surrounding qualia and consciousness, there are substantial philosophical disagreements about fundamental aspects of mentality. Although these disagreements do not by themselves impugn any particular (anti-physicalist) account, they provide some reason to doubt that we are dealing with straightforward ontological issues that can be resolved by appealing to a body of evidence about which everyone agrees. Furthermore, the debates tend to concern the very nature of the mental. More often than not, they involve specific assertions about mental phenomena. Given this, a successful anti-physicalism likely needs to provide some positive account of the mental, and this is precisely when the dilemma kicks in.

5.3. *Objection #3*

DaM does not provide support for physicalism.

Reply: We agree.²⁰ That was never the point. Indeed, we would go further and claim that DaM creates problems for various forms of physicalism. For instance, many won't work as a means of avoiding DaM because we still need to know what the mental entities are that supervene on, are realized by, or are grounded in the physical. Full-throated forms of eliminativism (Churchland, 1981; Bickle, 2003) may avoid this problem, but such views have their own challenges.

In addition, DaM undermines the popular strategy of defending physicalism by positing a No Fundamental Mentality constraint (Wilson 2006) or equating the physical with the non-mental (Spurrett and Papineau 1999; Montero 2009; Montero and Papineau 2005). Let's consider the constraint approach first. If we choose our current understanding of the mental, then any No Fundamental Mentality constraint will be trivially true simply because our current understanding of the mental is likely to be false. If we

²⁰ After all, physicalism still faces Hempel's original dilemma. For discussion and proposed solutions see Dove (2016); Elpidorou and Dove (2018); Hempel (1980); Ney (2008); Prelevic (2017); and van Fraassen (2002).

choose some future complete and ideal theory, then the proposed constraint is going to be empty. Similar problems face the stratagem of equating the physical with the non-mental. More generally, DaM further muddies the water with respect to the question of whether or not a complete physics contains fundamental mental entities.

5.4. *Objection #4*

Just as Hempel's Dilemma turns on an appeal to theories of the physical, DaM requires an appeal to theories of the mental. This is a problem for DaM because we do not have—or need—analogous theories of the mental. To the extent that a philosophical position avoids an appeal to theories of the mental, it avoids the dilemma.

Reply: Not surprisingly, we think that it is harder to avoid theoretical claims about the mental than this objection presupposes, but we maintain that the objection fails even if we leave this presupposition unchallenged. DaM concerns our lack of access to characterizations of the mental that are rich enough to settle important metaphysical questions. Avoiding theories of the mental does not get us out of this bind. We still have reason to think that characterizations based on our current understanding of the mental are insufficient and that we know too little about future characterizations to draw significant conclusions. Indeed, the situation appears to be worse than it is with regard to the physical. For instance, some supporters of physicalism have argued that we know enough about current physics to be confident that the posits of a complete and ideal *future* physical theory will exclude irreducible mental entities (e.g., Bokulich 2011; for reviews see Dove 2016; Ney 2008; Stoljar 2015). Yet, in the case of the mental, there is little indication that we know enough about what a future account of the mental would look like to offer a sufficiently rich characterization of the mental. A lack of access to theories of the mental would only exacerbate this problem.

In the end, the real impetus behind this objection would seem to be a conviction that our current understanding of the mental is rich enough to do the relevant philosophical work. We have already provided reasons to think that it is not. Whether or not one is convinced by our arguments on this front, it is clear that simply avoiding theories of the mental does not circumvent the dilemma.

5.5. *Objection #5*

Even if theories of the mental are relevant, most of our current theories are not about the mental itself but rather the metaphysical status of the mental with respect to the physical.²¹ In other words, our so-called theories of the mental are not theories about mental phenomena the way that physical theories are about physical phenomena. Because of this, the dilemma fails to emerge in the first place.

Reply: We have two responses to offer. First, while we acknowledge that there are theories of the mental that focus primarily on answering questions concerning the relationship of the mental to the physical, it is clear that not all of them are content to do only that. For instance, both proponents of naturalistic dualism (Chalmers 1996) and panpsychism (see various essays in Seager 2020) have sought to offer substantive, positive accounts of the mental.

Second, even if we were to grant that most theories of the mental are primarily interested in investigating the relationship of the mental to the physical, we maintain that it is difficult to answer these questions without answering fundamental questions about the nature of mental phenomena themselves. In order to see why this is the case, we need to consider the two horns of our dilemma.

Let's begin with the first horn of the dilemma. In our discussion, we highlighted the disagreement that exists with regard to fundamental questions concerning the nature of qualia. These questions were not limited to the relationship of qualia to the physical. Instead, they often concern important details about qualia themselves, addressing such issues as whether they exist at all, how they are introspectively revealed to us, whether they are simple or complex, what experiential modalities give rise to them, and what sort of creatures experience them. What is important to note isn't merely the existence of this disagreement, but also the fact that the manner in which we might resolve these theoretical disagreements has clear

²¹ We thank an anonymous reviewer for pointing out this possible objection. As they succinctly put it, theories of the mental "are not about the mental the way that fluid mechanics is about fluids, they are about the relationship between the mental and the physical."

implications regarding the metaphysical status of the mental and its relationship to the physical. After all, much of the recent debate regarding the metaphysics of consciousness seems to turn on whether the nature of qualia is somehow revealed to us through introspection.²²

Now consider the second horn of the dilemma. In outlining the problems with futurism, we did note that positions such as dualism, neutral monism, and panpsychism promise to be transformative. We also suggested that it would be hard to offer substantial versions of these positions without making significant claims about the nature of mental phenomena in and of themselves, but here we don't need this claim to defeat the objection. If the objector is right that contemporary theories of the mental merely address its relationship to the physical, then such theories would provide *no help at all* to futurism. They wouldn't even address the issues raised in the context of currentism.

6. Conclusion

In this essay we have argued that our understanding of the mental faces a similar (but not completely analogous) dilemma to the one facing our understanding of the physical. Our defense of this involved three steps. First, we outlined the reasons why this dilemma may have been overlooked or quickly dismissed. We argued that these reasons are insufficient and provided initial motivation for thinking that the dilemma might obtain. Second, we demonstrated that both horns of the dilemma are problematic: currentism with respect to mental is likely to be at least incomplete or inadequate, and futurism remains indeterminate. Third, we defended the dilemma against several deflationary objections. If we are right, then philosophers interested in metaphysical questions surrounding the mental need to take our dilemma into account.

²² See, e.g., Balog (2012); Chalmers (2007); Diaz-Leon (2014); Elpidorou (2015) and (2016); Goff (2011); Hill & McLaughlin (1999); Levin (2007); Levine (2007); Loar (1997); Nida-Rümelin (2006); Papineau (2002) and (2007); Schroer (2010), Stoljar (2005); Sturgeon (1994); Sundström (2011); Trogdon (2016).

It is instructive to compare our position to Stoljar's (2006) epistemic view on consciousness, namely, the view according to which the reason why consciousness appears to be other than physical is because we are ignorant of some non-experiential but experience-relevant truths. Unlike Stoljar we do not use our current ignorance about the nature of qualia or consciousness as a way to disarm traditional anti-physicalist objections. Nor do we insist that our ignorance is due to an elusive set of non-experiential but experience-relevant truths. To hold that the only truths about consciousness that escape us are non-experiential is to accept that our present understanding of (phenomenal) consciousness is more or less complete. DaM denies this assumption. Thus, if we are correct to insist that DaM is a problem, then our ignorance is larger than it is commonly assumed. The bad news is that we do not know as much as we think we do. The good news is that such an admission of ignorance opens up the possibility for new and exciting prospects on mentality in general and on consciousness in particular. A more complete understanding of the mental could render some of the pesky epistemic arguments against physicalism pseudo-problems. Or it could conversely show that physicalism is an unattainable position. So, while DaM may not resolve traditional philosophical puzzles, it may succeed in transforming them. At the very least, it is a call to action to seek a more philosophically and empirically robust account of the mental.

References

- Azzouni, Jody. 2004. "Theory, Observation, and Scientific Realism." *British Journal for the Philosophy of Science* 55 (3): 371–92.
<https://doi.org/10.1093/bjps/55.3.371>
- Bickle, John. 2003. *Philosophy and Neuroscience: A Ruthlessly Reductive Account*. Dordrecht: Kluwer Academic Publishers.
<https://doi.org/10.1007/978-94-010-0237-0>
- Balog, Katalin. 2012. "Acquaintance and the Mind-body Problem." In *New Perspectives on Type Identity: The Mental and the Physical*, edited by Simone Gozzano, and Christopher S. Hill, 16–42. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511687068.002>
- Block, Ned. 1980. "Troubles with Functionalism." In *Readings in the Philosophy of Psychology*, vol. 1, edited by Ned Block, 268–306. Cambridge: Harvard University Press.

- Block, Ned. 1990. "Inverted Earth." *Philosophical Perspectives* 4: 53–79.
<https://doi.org/10.2307/2214187>
- Block, Ned. 1995. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences* 18 (2): 227–47.
<https://doi.org/10.1017/S0140525X00038474>
- Bogen, Jim. 2016. "Empiricism and After." In *The Oxford Handbook of Philosophy of Science*, edited by Paul Humphreys, 779–95. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199368815.013.12>
- Bokulich, Peter. 2011. "Hempel's Dilemma and the Domain of Physics." *Analysis* 71(4): 646–51. <https://doi.org/10.1093/analys/anr087>
- Brewer, Bill. 2011. *Perception and its Objects*. Oxford: Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780199260256.001.0001>
- Campbell, John. 2002. *Reference and Consciousness*, Oxford: Clarendon Press.
<https://doi.org/10.1093/0199243816.001.0001>
- Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
- Chalmers, David J. 2003. "Consciousness and Its Place in Nature." In *The Blackwell Guide to Philosophy of Mind*, edited by Stephen P. Stich, and Ted A. Warfield, 102–42. Malden, MA: Blackwell Publishing. Revised version in Chalmers, 2010, 103–40. <https://doi.org/10.1002/9780470998762.ch5>
- Chalmers, David J. 2007. "Phenomenal Concepts and the Explanatory Gap." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter, and Sven Walter, 167–94. Oxford, UK: Oxford University Press.
<https://doi.org/10.1093/acprof:oso/9780195171655.003.0009>
- Chalmers, David J. 2010. *The Character of Consciousness*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195311105.001.0001>
- Chang, Hasok. 2005. "A Case for Old-fashioned Observability, and a Reconstructive Empiricism." *Philosophy of Science* 72 (5): 876–87.
<https://doi.org/10.1086/508116>
- Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*, Cambridge: Cambridge University Press.
<https://doi.org/10.1017/CBO9780511811937>
- Churchland, Paul M. 1995. *The Engine of Reason, the Seat of the Soul*. Cambridge, MA: MIT Press.
- Churchland, Paul M., and Patricia S. Churchland. 1992. "Intertheoretic Reduction: A neuroscientist's Field Guide." In *Neurophilosophy and Alzheimer's Disease*, edited by Yves Christen, and Patricia S. Churchland, 18–29. Berlin: Springer.
https://doi.org/10.1007/978-3-642-46759-2_3
- Churchland, Patricia S. 1986. *Neurophilosophy*. Cambridge, MA: MIT Press.

- Crane, Tim. 2001. "The Origins of Qualia." In *History of the Mind-Body Problem*, edited by Tim Crane, Sarah Patterson, 164–94. London: Routledge.
- Crane, Tim. 2007. "Knowledge of the Mind and Knowledge of the Brain." *3rd Annual Brain & Mind Lecture*, University of Copenhagen.
http://www.timcrane.com/uploads/2/5/2/4/25243881/knowledge_of_mind_and_knowledge_of_brain.pdf
- Crane, Tim, and D. Hugh Mellor. 1990. "There is No Question of Physicalism." *Mind* 90: 185–206.
- Diaz-Leon, Esa. 2014. "Do A Posteriori Physicalists Get Our Phenomenal Concepts wrong?" *Ratio* 27 (1): 1–16. <https://doi.org/10.1111/rati.12018>
- Dennett, Daniel C. 1991. *Consciousness Explained*. London: Little, Brown & Co.
- Dretske, Fred I. 1995. *Naturalizing the Mind*, Cambridge, MA: MIT Press, Bradford Books.
- Dove, Guy. 2018. "Redefining Physicalism." *Topoi* 37: 513–22.
<https://doi.org/10.1007/s11245-016-9405-0>
- Dowell, Janice L. 2006. "The Physical: Empirical, Not Metaphysical." *Philosophical Studies* 131 (1): 25–60. <https://doi.org/10.1007/s11098-006-6641-y>
- Duhem, Pierre Maurice Marie. 1914/1954. *The Aim and Structure of Physical Theory*. Translated by P. W. Wiener. Princeton, NJ: Princeton University Press.
- Elpidorou, Andreas. 2017. "Introduction: The Character of Physicalism." *Topoi* (37): 435–55. <https://doi.org/10.1007/s11245-017-9488-2>
- Elpidorou, Andreas. 2015. "Phenomenal Concepts." In *Oxford Bibliographies in Philosophy*, edited by Duncan Pritchard. New York: Oxford University Press.
<https://doi.org/10.1093/OBO/9780195396577-0254>
- Elpidorou, Andreas. 2016. "A Posteriori Physicalism and Introspection." *Pacific Philosophical Quarterly* 97 (4): 474–500. <https://doi.org/10.1111/papq.12068>
- Elpidorou, Andreas, and Guy Dove. 2018. *Consciousness and Physicalism: A Defense of a Research Program*. New York: Routledge.
<https://doi.org/10.4324/9781315682075>
- Feyerabend, Paul K. 1959/1985. "An Attempt at a Realistic Interpretation of Experience." In *Realism, Rationalism, and Scientific Method: Philosophical Papers I*, 17–36. Cambridge, UK: Cambridge University Press.
<https://doi.org/10.1017/CBO9781139171526.003>
- Flanagan, Owen J. 1992. *Consciousness Reconsidered*. Cambridge, MA: MIT Press.
- Gillett, Carl, and D. Gene Witmer. 2001. "A 'Physical' Need: Physicalism and the Via Negativa." *Analysis* 61 (272): 302–09. <https://doi.org/10.1111/1467-8284.00309>
- Goff, Philip. 2011. "A Posteriori Physicalists Get Our Phenomenal Concepts Wrong." *Australasian Journal of Philosophy* 89 (2): 191–209.
<https://doi.org/10.1080/00048401003649617>

- Hanson, Norwood Russell. 1958. *Patterns of Discovery*. Cambridge, UK: Cambridge University Press.
- Harman, Gilbert, 1990, "The Intrinsic Quality of Experience." In *Philosophical Perspectives*, vol. 4, edited by James Tomberlin, 31–52. Atascadero, CA: Ridgeview Publishing Company. <https://doi.org/10.2307/2214186>
- Hempel, Carl G. 1969. "Reduction: Ontological and Linguistic Facets." In *Philosophy, Science and Method: Essays in Honor of Ernest Nagel*, edited by Patrick Suppes, Sidney Morgenbesser, and Morton Gabriel White, 179–99. New York: St. Martin's Press.
- Hempel, Carl G. 1980. "Comments on Goodman's Ways of Worldmaking." *Synthese* 45: 139–99. <https://doi.org/10.1007/BF00413558>
- Hill, Christopher S., and Brian P. McLaughlin. 1999. "There Are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy". *Philosophy and Phenomenological Research* 59: 445–54. <https://doi.org/10.2307/2653682>
- Jackson, Frank. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (127): 127–36. <https://doi.org/10.2307/2960077>
- Jackson, Frank. 1986. "What Mary didn't know." *The Journal of Philosophy* 83 (5): 291–95. <https://doi.org/10.2307/2026143>
- Jackson, Frank. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/0198250614.001.0001>
- Keeley, Brian L. 2009. "The Early History of the Quale and Its Relation to the Senses." In *Routledge Companion to Philosophy of Psychology*, edited by Sarah Robins, John Symons, and Paco Calvo, 71–89. London, UK: Routledge.
- Kim, Jaegwon. 1998. *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.
- Kuhn, Thomas S. 1962/1996. *The Structure of Scientific Revolutions*, Chicago, IL: University of Chicago Press.
- Ladyman, James, and Don Ross. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199276196.001.0001>
- Levin, Janet. 2007. "What is a phenomenal concept?" In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter, and Sven Walter, 87–110. New York, NY: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195171655.003.0006>
- Levine, Joseph. 1983. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (4): 354–61. <https://doi.org/10.1111/j.1468-0114.1983.tb00207.x>
- Levine, Joseph. 2007. "Phenomenal Concepts and the Materialist Constraint." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on*

- Consciousness and Physicalism*, edited by Torin Alter, and Sven Walter, 81–108. New York, NY: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195171655.003.0008>
- Levine, Joseph, and Kelly Trogon. 2009. “The Modal Status of Materialism.” *Philosophical Studies* 145 (3): 351–362. <https://doi.org/10.1007/s11098-008-9235-z>
- Libet, Benjamin. 1985. “Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action.” *Behavioral and Brain Sciences* 8: 529–39. <https://doi.org/10.1017/S0140525X00044903>
- Loar, Brian. 1997. “Phenomenal states.” In *The Nature of Consciousness: Philosophical Debates*, edited by Ned Block, Owen Flanagan, and Güven Güzeldere, 597–616. Cambridge, MA: MIT Press.
- Lycan, William G. 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.
- Mayr, Ernst. 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge, MA: Harvard University Press.
- Melnyk, Andrew. 2003. *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511498817>
- Melnyk, Andrew. 1997. “How to Keep the ‘Physical’ in Physicalism.” *Journal of Philosophy* 94: 622–37. <https://doi.org/10.2307/2564597>
- Milner, David, and Mel Goodale. 2006. *The Visual Brain in Action*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198524724.001.0001>
- Montero, Barbara. 2001. “Post-Physicalism.” *Journal of Consciousness Studies* 8: 61–80.
- Montero, Barbara. 2009. “What is the Physical?” In *Oxford Handbook of the Philosophy of Mind*, edited by Brian McLaughlin, Ansgar Beckermann, and Sven Walter. Oxford: Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199262618.003.0010>
- Montero, Barbara, and David Papineau. 2005. “A Defense of the via Negativa Argument for Physicalism.” *Analysis* 65(3): 233–37. <http://www.jstor.org/stable/3329031>
- Nagel, Thomas. 1974. “What Is It Like to Be a Bat?” *The Philosophical Review* 83(4): 435–50. <https://doi.org/10.2307/2183914>
- Ney, Alyssa. 2008. “Defining Physicalism.” *Philosophy Compass* 3(5): 1033–48. <https://doi.org/10.1111/j.1747-9991.2008.00163.x>
- Nida-Rümelin, M. 2006. “Grasping Phenomenal Properties”. In *Phenomenal Concepts and Phenomenal Knowledge: New essays on Consciousness and Physicalism*, edited by Torin Alter, and Sven Walter, 309–38. Oxford, UK:

- Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195171655.003.0013>
- Papineau, David. 2001. "The Rise of Physicalism." In *Physicalism and Its Discontents*, edited by Carl Gillett, and Barry Loewer, 3–36. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511570797.002>
- Papineau, David. 2002. *Thinking About Consciousness*. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/0199243824.001.0001>
- Papineau, David. 2007. "Phenomenal and perceptual concepts." In *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*, edited by Torin Alter, and Sven Walter, 111–44. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195171655.003.0007>
- Peacocke, Christopher. 1983. *Sense and Content*. Oxford: Oxford University Press.
- Robb, David, and John Heil. 2014. "Mental Causation." In *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), edited by Edward N. Zalta., et al. <https://plato.stanford.edu/archives/spr2014/entries/mental-causation/>
- Seager, William E., ed. 2019. *The Routledge Handbook of Panpsychism*. New York, NY: Routledge. <https://doi.org/10.4324/9781315717708>
- Schroer, Robert. 2010. "Where's the Beef? Phenomenal Concepts as both Demonstrative and Substantial." *Australasian Journal of Philosophy* 88 (3): 505–22. <https://doi.org/10.1080/00048400903143861>
- Spurrett, David, and David Papineau. 1999. "A Note on the Completeness of 'Physics'." *Analysis* 59: 25–29. <https://doi.org/10.1093/analys/59.1.25>
- Stoljar, Daniel. 2001. "Two Conceptions of the Physical." *Philosophy and Phenomenological Research* 62 (2): 253–81. <https://doi.org/10.1111/j.1933-1592.2001.tb00056.x>
- Stoljar, Daniel. 2005. "Physicalism and Phenomenal Concepts." *Mind & Language* 20 (5): 469–95. <https://doi.org/10.1111/j.0268-1064.2005.00296.x>
- Stoljar, Daniel. 2006. *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*. Oxford, UK: Oxford University Press. <https://doi.org/10.1093/0195306589.001.0001>
- Stoljar, Daniel. 2015. "Physicalism." In *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), edited by Edward N. Zalta., et al. <https://plato.stanford.edu/archives/win2017/entries/physicalism/>
- Sturgeon, Scott. 1994. "The Epistemic View of Subjectivity." *The Journal of Philosophy* 91 (5): 221–35. <https://doi.org/10.2307/2940751>
- Sundström, Pär. 2011. "Phenomenal Concepts." *Philosophy Compass* 6 (4): 267–81. <https://doi.org/10.1111/j.1747-9991.2011.00384.x>
- Tiehen, Justin. 2018. "Physicalism." *Analysis*, 78 (3): 537–51. <https://doi.org/10.1093/analys/any037>

- Tiehen, Justin. 2016. "Physicalism Requires Functionalism: A New Formulation and Defense of the Via Negativa." *Philosophy and Phenomenological Research* 92 (2): 3–24. <https://doi.org/10.1111/phpr.12279>
- Trogdon, Kelly. 2016. "Revelation and physicalism." *Synthese*, 1–22. <https://doi.org/10.1007/s11229-016-1055-7>
- Tye, Michael. 2000. *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.
- Tye, Michael. 2017. "Qualia." In *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), edited by Edward N. Zalta., et al. <https://plato.stanford.edu/archives/win2017/entries/qualia/>
- Van Fraassen, Bas C. 2002. *The Empirical Stance*. New Haven: Yale University Press.
- Vicente, Agustín. 2011. "Current Physics and 'The Physical.'" *British Journal for the Philosophy of Science* 62 (2): 393–416. <https://doi.org/10.1093/bjps/axq033>
- Weiskrantz, L. 1986. *Blindsight: A Case Study and Implications*. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780198521921.001.0001>
- Wilson, Jessica. 2006. "On Characterizing the Physical." *Philosophical Studies* 131: 61–99. <https://www.jstor.org/stable/25471799>
- Witmer, D. Gene. 2016. "Physicality for Physicalists." *Topoi* (37): 457–72. <https://doi.org/10.1007/s11245-016-9415-y>
- Worley, Sara. 2006. "Physicalism and the Via Negativa." *Philosophical Studies* 131: 101–26. <https://doi.org/10.1007/s11098-005-5985-z>

Erratum: Normative Naturalism on Its Own Terms

Pekka Väyrynen*


Organon F 28 (3) 2021: 505–530. <https://doi.org/10.31577/orgf.2021.28302>

Published: 24 July 2021 / Updated: 22 November 2021

The original version of this article fails to credit the original source for a point made on p. 513: “Any natural language can have only countably many predicates, but natural properties might not be only countably many.” This sentence should have come with the following footnote: “Sturgeon credits to Richard Boyd the related point that there seem to be continuum many physical states of the world, and hence more physical properties than there are, even in the language of ideal physics, physical expressions to represent them” (Sturgeon 1985, 61). The author apologizes for the omission.

* University of Leeds

 <https://orcid.org/0000-0003-4066-8577>

 The School of Philosophy, Religion and History of Science, University of Leeds,
Leeds, LS2 9JT, United Kingdom

 p.vayrynen@leeds.ac.uk

© The Author. Journal compilation © The Editorial Board, *Organon F*.



This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International Public License (CC BY-NC 4.0).

Contents

EDITORIALS

Piotr Stalmaszczyk: <i>Preface</i>	1/2–8
Maciej Witek: <i>Preface</i>	2/270–281
Dan Zeman: <i>Introduction: “Value in Language”</i>	3/498–504

RESEARCH ARTICLES

José Ángel Gascón: <i>Why Did You Really Do It? Human Reasoning and Reasons for Action</i>	4/845–866
Stina Bäckström: <i>Must Expression Be Instrumental?</i>	2/282–302
Marina Bakalova: <i>The Epistemic Value of Music</i>	2/303–325
Yavuz Recep Başoğlu: <i>How Not to Argue about the Compatibility of Predictive Processing and λE Cognition</i>	4/777–801
Felix Bräuer: <i>Common Ground, Conversational Roles and Epistemic Injustice</i>	2/399–419
Bianca Cepollaro: <i>The Moral Status of the Reclamation of Slurs</i> ...	3/672–688
Eros Corazza – Christopher Genovesi: <i>On Anaphors Linked to Names Used Metaphorically</i>	1/253–268
Alice Damirjian: <i>Rethinking Slurs: A Case Against Neutral Counterparts and the Introduction of Referential Flexibility</i>	3/650–671
Alex Davies: <i>Faultless Disagreement Contextualism</i>	3/557–580
Guy Dove – Andreas Elpidorou: <i>A Dilemma about the Mental</i>	4/867–895
Juliana Faccio Lima: <i>How Can Millians Believe in Superheroes?</i> ...	1/135–167
Katharina Felka: <i>‘Boys Don’t Cry’ – An Ambiguous Statement?</i> ...	3/581–595
Nathan Hawkins: <i>Frege’s Equivalence Thesis and Reference Failure</i>	1/198–222
Zuzanna Jusińska: <i>Slur Reclamation – Polysemy, Echo, or Both?</i>	3/689–707

Natalia Karczevska: <i>Illocutionary Disagreement in Faultless Disagreement</i>	3/531–556
Miloš Kosterec: <i>Some Further Remarks on Hybrid View of Fictional Characters</i>	2/484–491
Marcin Lewiński: <i>Conclusions of Practical Argument: A Speech Act Analysis</i>	2/420–457
Chang Liu: <i>The Derogatory Force and the Offensiveness of Slurs</i> ..	3/626–649
Alba Moreno – Eduardo Pérez-Navarro: <i>Beyond the Conversation: The Pervasive Danger of Slurs</i>	3/708–725
Eleonora Orlando: <i>Fictional Names and Fictional Concepts: A Moderate Fictionalist Account</i>	1/107–134
Stefan Petkov: <i>The Degrees of Understanding and the Inferential Component of Understanding</i>	4/746–776
Stefano Predelli: <i>Fictional Tellers: A Radical Fictionalist Semantics for Fictional Discourse</i>	1/76–106
Stefano Predelli: <i>Unmentionables: Some Remarks on Taboo</i>	3/726–744
Louis Rouillé: <i>Anti-Realism about Fictional Names at Work: A New Theory for Metafictional Sentences</i>	1/223–252
Mark Sainsbury: <i>Fictional Names: Reference, Definiteness and Ontology</i>	1/44–59
Fiora Salis: <i>The Meanings of Fictional Names</i>	1/9–43
Merel Semeijn – Edward N. Zalta: <i>Revisiting the ‘Wrong Kind of Object’ Problem</i>	1/168–197
Andrés Soria-Ruiz: <i>Value and Scale: Some Observations and a Proposal</i>	3/596–625
Serdal Tümkaya: <i>A Novel Reading of Thomas Nagel’s “Challenge” to Physicalism</i>	4/802–818
Pekka Väyrynen: <i>Normative Naturalism on Its Own Terms</i>	3/505–530
Alberto Voltolini: <i>Real Authors and Fictional Agents (Fictional Narrators, Fictional Authors)</i>	1/60–75
Maciej Witek: <i>Self-Expression in Speech Acts</i>	2/326–359
Mateusz Włodarczyk: <i>Limitations of Non-Gricean Approaches to the Evolution of Human Communicative Abilities</i>	2/360–398

Kyoung-Eun Yang: <i>Do Kuhn's Cases of the Theory-Change from Newtonian to Einsteinian Physics Support His Incommensurability Thesis?</i>	2/458–483
Vojtěch Zachník: <i>Epistemic Foundations of Salience-Based Coordination</i>	4/819–844

BOOK REVIEWS

Jaroslav Peregrin: <i>Sanford Shieh: Necessity Lost: Modality and Logic in Early Analytic Philosophy, Vol. 1</i>	2/492–495
--	-----------

MISCELLANEA

Pekka Väyrynen: <i>Erratum: Normative Naturalism on Its Own Terms</i> ...	4/896
---	-------