The *Unconscious Mind Worry*: A Mechanistic-Explanatory Strategy

Beate Krickel (beate.krickel@tu-berlin.de)

Recent findings in different areas of psychology and cognitive science have brought the discussion of the unconscious mind back to center stage. However, the *unconscious mind worry* remains: What renders unconscious phenomena mental? In the present paper, I will suggest a new strategy for answering this question. This strategy rests on the idea that categorizing unconscious phenomena as "mental" should come out as *scientifically useful* relative to the explanatory goals of unconscious mind research. I will argue that this is the case if by categorizing an unconscious phenomenon as "mental" one picks out *explanatorily relevant similarities* between that phenomenon and a corresponding paradigmatically mental phenomenon, i.e., a conscious one. Explanatory relevance is spelled out in terms of the mechanistic norms of scientific explanation.

1 Introduction

Most contemporary philosophers, psychologists, and cognitive scientists agree that there is an unconscious mind. Motivated by recent empirical findings in different areas of psychology, it is argued that vision can occur unconsciously (Prinz 2015; Phillips and Block 2016; Peters et al. 2017; Berger and Nanay 2016), that our behavior is often influenced by unconscious attitudes, preferences, and evaluations (Mandelbaum 2016; Gawronski, Hofmann, and Wilbur 2006; Banaji, Bhaskar, and Brownstein 2015), that decision-making (Dijksterhuis and Nordgren 2006) and learning can occur unconsciously (Pothos 2007), that even emotions and affective processes can be unconscious (Smith and Lane 2016), and that thought and reasoning occur unconsciously when cognitive inconsistencies need to be resolved (Carruthers 2015; Quilty-Dunn and Mandelbaum 2018). Based on these findings, the psychologists John Bargh and Ezequiel Morsella conclude that "[i]n nature, the 'unconscious mind' is the rule, not the

exception" (Bargh and Morsella 2008, 78). Similarly, in a review article, Jacob Berger observes that "[t]here is growing consensus not only in psychology and neuroscience but also in philosophy that mental states can and often do occur outside of consciousness" (Berger 2014, 392).

Claims to the existence of an unconscious mind are not truisms. They go beyond the uncontroversial claim that we are not conscious of some factors that influence our thoughts and behaviors. E.g., clearly, we are not aware of everything that is happening in our brains. We are not aware of the firings of the billions of neurons that make up our brain. We are not even aware of all neural and bodily goings-on that process information. In the absence of our awareness, a huge amount of neural and cellular activity is involved in the processing, transmission, and evaluation of incoming signals. Furthermore, there are various physical and biological factors that influence our behaviors of which we are not conscious. Gravity, air pressure, etc. constrain our bodily movements; physiological processes, such as hunger, or tiredness often influence our behavior; and biological factors, such as hormones, influence our social interactions. Although these factors and their influences are (usually) not conscious to us, they are not taken to provide evidence for an unconscious mind.

What distinguishes claims to the existence of an unconscious mind from the trivial cases? What justifies talking about unconscious *mental* influences on thought and behavior over and above merely influences that are not conscious to us? This is what I call the *unconscious mind worry*. The worry has a *methodological* reading as well as a *conceptual* one. The methodological version concerns the question of how to empirically determine that an unconscious phenomenon is mental. The conceptual version concerns the conditions under which we are justified in ascribing the predicate 'mental' to unconscious processes.

In so far as the unconscious mind worry is addressed at all, the contemporary empirical and philosophical literature on the unconscious mind mainly discusses the methodological reading

of it, i.e., the question of whether the empirical tests and experimental designs employed to investigate the unconscious mind are indeed adequate. This focus is problematic. It neglects the fact that we must have a clear understanding of what we are looking for before we can discuss whether a test or experimental design is adequate for detecting it. Disagreements about methodological issues can be fruitfully discussed only based on agreements regarding what the methodology is supposed to deliver. Thus, to clarify under which conditions ascribing mentality is justified, we first must address the *conceptual unconscious mind worry*.

The paper proceeds as follows: in Section 2, I summarize the state-of-the-art concerning empirical research on two prominent examples of putative unconscious mental phenomena (unconscious perception and unconscious bias). In Section 3, I will summarize general attempts to address the conceptual unconscious mind worry that can be found in the philosophical and psychological literature. It will turn out that none of them is successful. In Section 4, I will develop a new strategy for addressing the conceptual unconscious mind worry. In the Conclusion, I will summarize the results and briefly discuss the consequences of my account regarding the *methodological unconscious mind worry*.

2 Examples of Putatively Unconscious Mental Phenomena & the Methodological Unconscious Mind Worry

2.1 Unconscious Vision?

One (putative) unconscious mental phenomenon that currently receives a lot of attention in the philosophical and psychological literature is unconscious vision. The most prominent case of putatively unconscious vision is so-called *blindsight* in patients with lesions in specific areas of the visual cortex (V1). These patients report not seeing anything in the contralateral visual field. Still, in forced choice visual detection tasks, they can detect stimuli in this part of their visual field with great accuracy (Weiskrantz 1995). Whether blindsight in patients really provides evidence for unconscious vision remains controversial. One reason is that it cannot be excluded

that the patients' brain damages lead to a distortion of the capacity to report visual stimuli (Persuh 2018). This objection can be avoided by focusing on non-pathological cases. (Putative) unconscious vision could be observed in healthy subjects by using different paradigms, such as binocular rivalry (Yang, Zald, and Blake 2007).

Binocular rivalry is an experimental paradigm in which each eye is presented with a different picture with the results that the two pictures alternately fade into and out of consciousness. Binocular rivalry is used in different variants to investigate putative unconscious vision (see Phillips and Block (2016) for an overview). When presenting a so-called "Mondrian" (patterns of colored or black and white rectangles) to one eye, the conscious perception of the image presented to the other eye can be suppressed for minutes – subjects will report that they do not see the other image. However, if the other eye is presented with a picture that depicts, for example, nudes it will attract or repel the subject's spatial attention depending on the subject's gender and sexual preferences (Jiang et al. 2006)[1]. This was established by having subjects categorize Gabor patches that were presented after the Mondrian/nude trial in either the location of the nude or the location of the Mondrian. Subjects were asked to indicate the orientation of the Gabor patch. One result was that male subjects were more accurate in categorizing the Gabor patches if they appeared at the location of the picture of the nude if the nude was female; they were less accurate when the Gabor patches were at the location of a previously presented nude male. Block (2016) argues that these findings provide evidence for unconscious vision: first, subjects report not seeing anything but the Mondrian; hence whatever is happening is not conscious to the subjects. Second, the fact that pictures of nudes attract or repel the subjects' attention depending on their gender and sexual orientation can only be explained by the assumption that they are perceived. Hence, there is unconscious vision.

[1] In the study conducted by Jing et al, two Mondrians and a fixation cross were presented to one eye, and one Mondrian, a fixation cross, and a picture of a nude was presented to the other eye.

However, not all researchers are convinced that the results of these studies provide evidence for unconscious vision (Newell and Shanks 2014; Phillips 2018). One criticism, which Phillips (2018) calls the *problem of the criterion*, targets the most-common method used to establish that subjects are not conscious of the stimuli: subjective reports of absence of conscious vision. It is objected that such reports are compatible with degraded conscious vision (Peters et al. 2017). A second objection, which Phillips (2018) calls the *problem of attribution*, holds that it is unclear whether the empirical methods justify the postulation of unconscious *vision*, i.e., of a mental phenomenon. According to Phillips (following Burge (2010)) vision is an individual-level phenomenon. The problem of attribution arises for empirical studies that "fail to show that the states they implicate are attributable to the individual" (Phillips 2018, 481). Evidence for whether a process is attributable to the individual would consist in showing that the process is "available to central coordinating agency" (Phillips 2018, 494). Studies using unconscious visual primes are confronted with this problem since the primes generate representations that are not available to central coordinating agency, and thus it is unclear why these representations should be attributable to the individual. Rather, after the presentation of the prime the perceptual system "is more 'fluent'" in processing further matching stimuli (Phillips 2018, 494). Phillips sees the same problem for the nude study: "[W]hy think that stimulus-driven, reflex-like attentional responses count as manifestations of central agency, and so witness individual-level perception?" (Phillips 2018, 495).

The *problem of the criterion* and the *problem of attribution* primarily concern methodological issues. Still, both presuppose conceptual assumptions. Phillips's formulation of the problem of attribution clearly presupposes an answer to the conceptual unconscious mind worry: for a state or process to be mental it must be attributable to the individual – which is the case if it is available to central coordinating agency. However, it is unclear whether a view of

the mental in terms of attributability to the individual is adequate. I will discuss this in more detail in Section 3.

2.2 Unconscious Attitudes?

A further example of a putatively unconscious mental phenomenon are implicit attitudes, or 'implicit biases'. Generally, attitudes are taken to be mental states that correspond to negative, positive, or neutral evaluations of different items. A big part of implicit attitude research focusses on social attitudes, i.e., people's evaluation of different social groups such as African Americans or women. One crucial finding is that subjects who report no negative evaluations of specific social groups, in implicit tests (see below) show behaviors that suggest a negative attitude towards these groups. One standard conclusion is that implicit attitudes, thus, are unconscious (Banaji, Bhaskar, and Brownstein 2015; Saul 2013).

Different methods have been suggested to detect implicit attitudes including masked affective priming (Degner et al. 2007), introspecting affective reactions (Hahn et al. 2014), the implicit associations test (IAT) (Greenwald, Mcghee, and Schwartz 1998) and explicit judgment tests, e.g., where subjects were asked to evaluate different made-up CVs by male and female applicants and decide who should be hired (Uhlmann and Cohen 2005). Many psychologists as well as philosophers interpret the results of such studies as evidence for an unconscious mind. For example Mandelbaum (2016) argues that

> [e]xamining the workings of implicit bias can illuminate a host of foundational issues in cognitive science, such as the entities that populate the unconscious mind, and how rationally responsive unconscious thought can be. (Mandelbaum 2016, 629)

The *methodological mind worry* arises because it is unclear how these different methods could in principle measure the same construct: introspecting affective reactions is adequate for detecting implicit attitudes only if implicit attitudes are affective states. However, the IAT is

usually interpreted as detecting associations between concepts (Greenwald, Mcghee, and Schwartz 1998). Other authors argue that implicit attitudes cannot be associations but must be propositionally structured beliefs (Mandelbaum 2016).

So far, no agreement has been reached about what kinds of entities implicit attitudes are supposed to be (Feest 2020). However, the unconscious mind worry concerns an even more basic question: given the empirical evidence, are we justified in assuming that implicit attitudes are states *of the mind* at all? There is evidence, for example, that IAT scores are influenced by physiological factors such as sleep, age, and hunger. This could indicate that, similar to Phillips's criticism presented in the previous section, the IAT simply measures the "fluency" of the brain in processing information rather than any mental phenomenon. A further explanation that would speak against categorizing implicit attitudes as mental states is that they are "truly implicit" (Johnson 2019). Contents are truly implicit if they are not represented in the system but a consequence of the specific architecture and transformation rules between contents. This discussion shows that unless the *conceptual* unconscious mind worry is solved, the discussion of whether unconscious bias research provides evidence for an unconscious mind is likely to be unproductive.

3 The *Conceptual Unconscious Mind Worry* & the Mark of the Mental

The conceptual unconscious mind worry is related to the philosophical search for the "mark of the mental". However, while the search for the mark of the mental is usually understood as the search for a necessary and sufficient condition for being mental, the conceptual unconscious mind worry does not require a solution in terms of a definition. Still, any proposal for the mark of the mental may be a solution for the conceptual mind worry. One feature that has been discussed as the mark of the mental is *consciousness*. Obviously, adopting this view would amount to a rejection of the unconscious mind. To not exclude the possibility of the unconscious mind from the get-go, we should look for an alternative.

One prominent proposal that is close to the first one while allowing unconscious mental phenomena is Searle's (1992) *connection principle*. According to this principle, unconscious phenomena are mental if they are *potentially conscious*. Against this proposal, it has been objected that the notion of *potentially* conscious remains "obscure" (Berger 2014, 399) as there is no clear account of what "potentially" means in this context. The approach that I am going to defend in Section 4 agrees with Searle's account that what renders the unconscious mental is its connection to conscious phenomena, i.e., that unconscious mental states or processes are mental because they are in some sense "potentially" conscious. My account diverges from Searle's in that the core argument is based on explanatory relevance of the term "mental" rather than any purported connection between intentionality and phenomenal consciousness. Furthermore, my account is epistemic or semantic – I clarify when it is scientifically useful to call an unconscious phenomenon "mental" rather than making any metaphysical claims about the nature of the unconscious. Still, my account can be interpreted as clarifying what it could mean for an unconscious mental state or process to be "potentially" conscious.

The second candidate for the mark of the mental is *intentionality*. Mental states, in contrast to physical states, are *about* something, sometimes even about non-existing objects. One standard way to make sense of intentionality is in terms of representations. This view is compatible with the existence of an unconscious mind. Unconscious vision, unconscious attitudes, etc. could be mental in the sense of having representational content. However, whether having representational content can possibly serve as the mark of the mental has been questioned (e.g., Berger (2014)). The main worry in the present context is that having representational content may not be sufficient for addressing the unconscious mind worry. Many philosophers and cognitive scientists hold that, for example, neural states, such as the "topographic map of the visual environment contained in primary visual cortex" are representational (Thomson and Piccinini 2018, 200). However, it is at least unclear whether

these states qualify as *mental* states. For example, Morgan and Piccinini (2018) conclude their paper on representations in cognitive neuroscience:

> A full neurocognitive explanation of intentionality will span mechanisms nested across multiple levels of organization in the nervous system and draw from ongoing developments in cognitive neuroscience. And there are other questions to keep philosophers busy: In virtue of what is mental content mental? Is mental content a distinct kind of content, or is it just the content had by mental states, which might in principle be had by other representational vehicles? (Morgan and Piccinini 2018, 136)

In the present context, this problem becomes even more pressing. The *unconscious mind worry* requires us to provide a criterion that distinguishes between the truism that we are not conscious of all influences on our thought and behavior and a more interesting unconscious mind thesis. However, whatever is going on in primary visual cortex, it is a truism that we are not conscious of what is going on there. Thus, the fact that we are not conscious of what is going on in the primary visual cortex should not come out as surprising and these processes—representational or not—thus should not be labeled "mental". Otherwise, it would remain unclear how the claim to the existence of an unconscious mind is supposed to be more than a truism.

Another proposal for distinguishing the mental from the non-mental is in terms of the personal/sub-personal distinction. While this distinction has originally been introduced by Dennett to denote different styles of explanation, it is commonly interpreted as an ontic distinction: there is a personal level – the level of mental states, and there is a sub-personal level – the level of the neuronal, biological, and physical mechanisms that may be involved in realizing the states of the personal level. The most crucial problem for an account in terms of the personal/sub-personal distinction is that the latter is usually spelled out in terms of the conscious/unconscious distinction (Drayson 2012, 15; Carter and Rupert 2020): processes at the personal level are conscious, whereas those at the sub-personal level are unconscious. Thus,

the possibility of unconscious mental phenomena is conceptually excluded. However, Burge's/Phillips's criterion for distinguishing the personal from the sub-personal introduced in Section 2.1 does not seem to fall prey of this problem. Following Burge. Phillips argues that a sufficient condition for being a personal-level phenomenon is to be "accessible to central agency". This kind of accessibility is supposed to be weaker than conscious accessibility (Phillips 2018, p. 493). Still, one problem is that it remains unclear how exactly Phillips wants us to understand the notion of "central coordinating agency" as there are different possible interpretations – and, as Shepherd and Mylopoulos (2021) argue, none of these seem to support Phillips's view. Furthermore, other objections that have been put forward against the personal/sub-personal distinction seem to apply to Burge's/Phillip's account as well: it has been argued that the identification of the personal level with the level of the mental is simply false (Figdor 2018; Dennett 2007). Again others have uttered general concerns about the fruitfulness of this distinction for cognitive science (Rupert 2018). While the last word may not have been spoken in this regard, these problems provide enough motivation to look for an alternative way for answering the conceptual unconscious mind worry.

A proposal for characterizing the mental that has explicitly been suggested by psychologists investigating the unconscious mind (Hassin 2013) is in terms of a functional characterization of mental capacities. Hassin argues for what he calls the "Yes It Can" principle: "Unconscious processes can carry out every fundamental high-level function that conscious processes can perform" (Hassin 2013, 195). Turning this into an answer to the conceptual unconscious mind worry, one could specify the functional role of the capacities that psychology and cognitive science take to be mental (or "cognitive" for that matter) and argue that unconscious phenomena are mental if they have the same functional role as their conscious counterpart (Hassin 2013; Berger 2014).

However, it is unclear how exactly to understand this proposal. On a fine-grained individuation of functional roles, unconscious states and processes do clearly not have the same functional roles as conscious ones. For example, the behaviors of subjects triggered by unconscious vision clearly differ from the behaviors of subjects that consciously see a stimulus. For example, they will be more accurate in detecting the stimulus, they may report seeing the stimulus, they may justify actions by referring to having seen the stimulus, and so on. The same applies to implicit attitudes. For one, "scores on implicit measures appear to be more temporally unstable than individuals' scores on corresponding explicit measures" (Brownstein, Madva, and Gawronski 2019, 6). Additionally, someone with a conscious attitude, will use this in their conscious reasoning, justification, imagination, and so on. Therefore, on a fine-grained reading of 'functional role', conscious and unconscious states and processes clearly do not have "virtually the same roles in our mental lives" (Berger 2014, 394). However, on a more coarse-grained individuation of the functional roles, e.g., 'being caused by external or internal factors and potentially influencing behavior and thought', the unconscious mind hypothesis turns out to be trivial as there are many cases in which an external or internal stimulus influences our behavior without consciousness being involved. One strategy may be to simply find the right-level of grain for individuating the functional roles that characterize mental capacities. The approach that I am going to defend in the next section can be understood as specifying the right level of grain. However, my approach goes beyond the functional role account as the crucial criterion for unconscious mentality, according to my proposal, is explanatory relevance rather than functional similarity.

4 A New Strategy for Addressing the *Conceptual Unconscious Mind Worry*

To address the *conceptual unconscious mind worry*, I want to propose a new strategy. This strategy starts from two assumptions: first, there is one sufficient condition for mentality, i.e., *consciousness*. Second, to determine the sense in which unconscious mental phenomena are

mental despite being unconscious, we must ask: *What is the scientific usefulness of categorizing certain unconscious phenomena as "mental", and others not?* If we want to find out whether, say, unconscious perception and unconscious attitudes are mental phenomena, we must analyze the general scientific reasons for ascribing mentality to unconscious phenomena, and find out whether these reasons apply to unconscious perception and unconscious attitudes. Note that my strategy is not to describe how psychologists and cognitive scientists de facto justify their mind ascriptions. On the one hand, they usually do not address this issue explicitly (for an exception see Section 3). On the other hand, they might not have any systematic and/or scientific reasons for talking about unconscious mental phenomena as opposed to merely unconscious phenomena. Thus, my project is a normative one: I want to analyze the implicit patterns of unconscious mind ascriptions, systematize them, and derive a proposal for how scientists (and philosophers) should use unconscious mind terminology to be consistent and to employ it in a scientifically useful way.

4.1 The Inference to the Unconscious Mind

Why do scientists label some unconscious phenomena "mental" and others not? To answer this question, it is helpful to take a closer look at the justification underlying the postulation of unconscious mental phenomena in general. The arguments for the existence of unconscious vision and unconscious attitudes as well as all other claims to the existence of an unconscious mind based on empirical evidence take the form of an *inference to the best explanation*:

Unconscious vision

1'. Subjects correctly react to visual stimuli above chance.

2'. Subjects do not consciously see the stimuli.

3'. The best explanation for how subjects can adequately react to stimuli despite not consciously seeing them is that they see them unconsciously.

4'. The explanation in terms of unconscious vision is true, and unconscious vision is real.

Unconscious attitudes

1". Subjects show discriminating behaviors towards women.

2". Subjects are not conscious of having any negative attitudes towards women.

3". The best explanation for why subjects discriminate against women despite not being conscious of any negative attitudes is that the attitudes are unconscious.

4". The explanation in terms of unconscious attitudes is true, and unconscious attitudes are real.

Generalizing these two examples, I will call the following argument the *Inference to the Unconscious Mind (IUM)*:

The Inference to the Unconscious Mind (IUM)

1. *Behavioral Observation*: Subjects show a certain behavior B.
2. *Unconsciousness*: Subjects are not conscious of a mental cause M of B.
3. *Best Explanation*: The best explanation for the conjunction of 1 and 2 is that B is due to M and that M is unconscious.

---

4. *Reality of Unconscious Mind:* The explanation in terms of unconscious mental causes is true, and the unconscious mental cause is real.

As indicated in Section 2, a lot of discussion in the philosophical and psychological literature targets premise 2. However, for the purposes of this paper, I will set these worries aside and assume the truth of premise 2. For the new strategy for characterizing the mental, premise 3 is crucial. This premise states that assuming that there is a mental cause that is unconscious provides an *explanation* of the behavior at issue (I will leave the discussion of the claim that these explanations are the *best* explanations for another paper). Two observations can be made: first, unconscious mental phenomena are postulated primarily as *explanatory posits*. Second, the explanatory power of the unconscious mental phenomenon seems to stem from its being mental in the same way as a corresponding conscious cause would have been. Thus, one implicit

assumption seems to be that unconscious mental phenomena are just like the corresponding conscious phenomena *minus consciousness*, and that it is this similarity that does the explanatory trick. My strategy makes use of this insight. If we find out how unconscious phenomena are similar to conscious phenomena in an explanatorily relevant way, we have determined what "mental" is supposed to grasp in premise 3. Thereby, we have identified the reasons for labelling some processes "mental" and others not: only the former are relevantly similar to corresponding conscious phenomena.

In which way are the explanatory roles of conscious and unconscious mental phenomena alike? And how does the explanatory role of non-mental unconscious phenomena differ? In a nutshell, I will argue that categorizing an unconscious phenomenon as "mental" is accurate if (a) its underlying mechanism produces and explains a behavior that is similar to a behavior produced and explained by a corresponding conscious phenomenon, and (b) the mechanism underlying the unconscious phenomenon shares explanatorily relevant components with the mechanism underlying the corresponding conscious phenomenon. To defend this view, I must show (i) that explanations invoking unconscious mental phenomena and those invoking corresponding conscious phenomena are *mechanistic explanations*; (ii) that they *explain similar behaviors*; and (iii) that the underlying mechanisms share explanatorily relevant components and how *explanatory relevance* is determined. I will elaborate on each of these points in the next sub-sections.

4.2 Mechanistic Explanation & the Unconscious Mind

What type of explanation are unconscious mental explanations supposed to deliver in scientific contexts? The general justification for the postulation of unconscious mental phenomena, as described in the IUM, is that people are not conscious of mental *causes* and that the best explanation for why they still show the respective *behavior* is that these mental causes are

unconscious. Researchers want to understand *how* the observed behavior comes about given that it cannot be due to a conscious perceptual state. They are interested in the *causal mechanism* that is responsible for the behavior. Hence, mental explanations, in the relevant context, are supposed to provide *mechanistic explanations* of behaviors.

Mechanistic explanations have become a core topic in contemporary philosophy of science. One common assumption of the so-called new mechanists is that *a mechanistic explanation of a phenomenon P consists of a description of a mechanism that is responsible for P*. "Responsible" in this context is usually understood in terms of either causation or constitution (Craver 2007). In causal/etiological mechanistic explanations, the mechanism that explains the phenomenon causes the phenomenon and, thus, temporally precedes it. In constitutive mechanistic explanations, the mechanism and the phenomenon occur in the same space time region and are mutually dependent.

Mechanisms are taken to be collections of entities and their activities in a certain spatial, temporal, causal, and hierarchical organization (Craver 2007; Glennan 2017; Illari and Williamson 2012; Bechtel and Abrahamsen 2005). Note that mechanistic components (i.e., entities and activities) can be described in different terminologies and with different grades of detail. They may, for example, be described in neurobiological terms (e.g., "neuron", "axon", "diffusing") or in functional terms (i.e., "inhibitor", "store") or filler terms (e.g., "activating", "inhibiting"). Explanations in terms of functional or filler terms are often described as "mechanism sketches" that still count as mechanistic explanations but need to be mapped onto a neurobiological/physical structure in order to be valid mechanistic explanations (Piccinini and Craver 2011).

Mechanisms are individuated relative to the phenomena they are responsible for. That is, the components of a mechanism for P are those entities and activities (in a certain organization) are all relevant for the production, and thus, explanation of P and all entities and activities (and

organizational aspects) that are relevant for the production and explanation of P are components of the mechanism for P. Which entities, activities, and organizational features are taken to be relevant for the production and explanation of P depends on whether they are causally relevant (in etiological mechanistic explanations) or constitutively relevant (in constitutive mechanistic explanations) for the phenomenon. According to a prominent proposal (Craver 2007), causal as well as constitutive relevance are determined by means of intervention-experiments. There is an on-going discussion of how to best understand this proposal (see, e.g., Kästner and Andersen (2018) for an overview). Luckily, the details of this discussion are not relevant here. The important aspect, here, is that philosophical discussions of mechanism discovery and individuation are closely tied to actual scientific practice – which is captured by the idea that the former rely on intervention-experiments. Thus, prima facie, the mechanistic approach to scientific explanation does not only capture the general idea of unconscious mind research (explaining how behaviors are brought about) but can also be applied to the search for mechanisms in unconscious mind research.

Are mechanistic explanations invoking unconscious mental phenomena etiological or constitutive mechanistic explanations? The answer to this question depends on what exactly those explanations explain. If the explanandum is, for example, "How does unconscious vision work?", the explanation would refer to a mechanism that constitutes unconscious vision. However, while this how-question may be a question that researchers want to answer, the explananda relevant in the present context are more fine-grained and relative to specific experimental setups. Furthermore, what is at issue here is the type of explanation that motivates the postulation of an unconscious mental phenomenon in the first place (see IUM in Section 4.1). In the context of the IUM, unconscious vision is the explanans, not the explanandum. Mechanistic explanations invoking unconscious mental phenomena take the form of "P happened because of unconscious vision/attitudes/…" and express that there is a mechanism

that is responsible for P that does not lead to reportability in the relevant sense but still is a mental mechanism (an instance of vision, believing, or the like). Thus, in the mechanistic context, "unconscious vision", "unconscious attitude" etc. can be understood as a filler-terms to be filled-in with details about the mechanism for P where this mechanism occurs in the absence of consciousness.

In the following section, I will argue that the explananda of explanations invoking unconscious mental phenomena are contrastive questions concerning behaviors observed in experiments. For example: Why could the patient accurately react to the stimuli rather than not? A mechanistic explanation of this contrastive explanandum picks out specific components of the mechanism underlying the patient's behavior that are causally relevant for the contrast referred to in the question. These are causal explanations rather than constitutive ones because whatever happened in the brain that led to the patient accurate reaction happened before the reaction.

4.3 The Explananda of the Unconscious Mind

What exactly are the explananda of explanations invoking unconscious mental phenomena? On the assumption that my reasoning in the previous section is correct, their explananda should be of the type that can principally be mechanistically explained. In contemporary philosophy of science, it is commonly assumed that what is to be explained by mechanistic explanations are phenomena that are picked out by how-does-x-work-questions (Craver 2007). These, in turn, are answered by answering a set of contrastive why-questions (Craver and Kaplan 2020). For example, in order to explain how the action potential is transmitted along the axon, one has to answer a bunch of contrastive why-questions such as "Why does the potential peak at 40mV rather than at a higher value?", "Why does the axon membrane hyperpolarize rather than go

back to resting potential?", "Why is the potential transmitted only in the direction of the axon terminal rather than back towards the soma?" etc.

What contrastive questions do unconscious mental phenomena answer? Generally, (conscious) mental phenomena are supposed to explain how a certain behavior is brought about by explaining *why a certain behavior occurred rather than some other behavior (or no behavior at all)* ("Why does B occur rather than not?"). For example, conscious perceptual states explain why a subject can react to visual stimuli with only a few mistakes rather than with many mistakes. Explicit negative attitudes towards Black people explain why 21.5% of a sample of undergraduates were neutral or agreed with the statement "Avoiding interactions with Blacks is important to my self-concept" (Axt 2018, 8). Postulating *unconscious* mental phenomena serves the same explanatory purpose. Take pathological blindsight: the ascription of an unconscious perceptual state explains why the behavioral reactions towards visual stimuli of blindsight subjects are quite accurate rather than not accurate at all. Furthermore, unconscious perceptual states explain why, in the binocular rivalry experiment described in Section 2.1, subjects' reactions to Gabor patches that are presented at the location of the sexually preferred nude are more accurate compared to Gabor patches that were presented at the location of the Mondrian - rather than reaction times being equal in both cases. And unconscious attitudes explain why self-reported non-racist people show discriminating behavior against members of certain social groups rather than treating members of different social groups equally in line with what they report. In a nutshell: Conscious as well as unconscious mental phenomena are postulated as parts of mechanisms that are supposed to explain behavioral contrasts of the form "Why did B occur rather than not-B?".

However, the similarities between scientific explanations invoking conscious phenomena and those invoking unconscious mental phenomena do not end here. The behavior that is supposed to be explained by conscious phenomena and the behavior that is supposed to be

explained by unconscious mental phenomena are of the same type – on not too fine-grained reading. This is the insight that motivates the functional approach to individuating mental capacities described in Section 3. The behavioral contrast in both cases is between a type of behavior that shows some kind of accuracy, stability, correctness, etc. compared to absence of that particular behavior. For example, conscious and unconscious vision are supposed to explain more or less accurate reactions towards visual stimuli in contrast to random reactions. The same holds for attitudes/preferences: Conscious and unconscious attitudes explain why subjects show more or less stable preferences towards something over something else rather than treating both equally (see Section 3). Someone who has the conscious attitude that women should not be police chiefs will very consistently reject female applicants for the job of a police chief; someone who is implicitly biased against women will be likely to reject female applicants but may be less consistently do so as their decision may be influenced by, for example, contextual factors.

To summarize: I have argued that explanations invoking unconscious mental phenomena are similar to explanations invoking conscious phenomena in the following ways: first, they both are supposed to deliver mechanistic explanations of how a behavior of type B is brought about by answering a question of the form "Why did B occur rather than not?". Second, the behavior B that is to be explained by the unconscious mental phenomenon is of the same type (not too fine-grainedly individuated) as the behavior that is to be explained by the corresponding conscious phenomenon. In the following section, I will explain how conscious and unconscious mental phenomena explain similar behavioral contrasts in a similar way – and thereby provide a novel approach to answering the conceptual unconscious mind worry in line with the general strategy depicted in Section 4.1.

4.4 Explanatory Relevance & The Unconscious Mind

In this section I will argue that scientific explanations invoking unconscious mental phenomena explain their explananda in a relevantly similar way as conscious phenomena explain their explananda – and that this is the crucial insight for addressing the *unconscious mind worry*. More specifically, I will argue that the categorization of an unconscious phenomenon as "mental" is adequate if the mechanism that is responsible for the behavior B* of type B when consciousness is absent and the mechanism that is responsible for the behavior B** of type B when consciousness is present share components that are relevant for explaining the contrast "Why B rather than not-B?". To get to this, I will first explain how to think of explanatory relevance in the mechanistic framework.

According to a recent mechanistic proposal (Kohár and Krickel 2021), in order to answer contrastive why-questions of the form "Why P rather than P'?" we have to look at the different mechanisms that would produce phenomenon P' and compare them to the actual mechanism that is responsible for phenomenon P. More specifically, we must compare the actual mechanism for P to the *maximally similar* mechanism for P' and list all the differences between the two mechanisms *that are shared by all possible mechanisms for P'*[2]. We thereby list all those components of the mechanism for P whose absence would be necessary for P' to arise, and those components that are lacking in the mechanism for P whose presence would be necessary for P' to arise. All items of this list can be called the *difference-makers for P vs. P'*.

To illustrate this idea, take the example of the action potential again. If one wants to explain why the action potential is transmitted in the direction of the axon terminal rather than back to the cell soma, one has to compare what is happening in the actual case to what would have to be the case in order for the potential to travel backwards and list all those elements that are

[2] For the sake of clarity, I will ignore this latter condition in the following. In the original approach, this condition is introduced to deal with multiple realization, i.e., cases where there are many different mechanisms that might lead to the contrast phenomenon.

present in the actual case that prevent the contrast phenomenon to be the case, and those elements that are lacking in the actual case that would be necessary for the contrast phenomenon to occur. Thereby, one reaches an answer such as: "… because, in the actual mechanism, the sodium potassium ion pump is still active after the potential reaches -70mv and potassium ions continue to move out of the axon rather than, as in the most similar possible mechanism for the action potential's traveling backwards, the sodium potassium ion pump becoming inactive when the potential reaches -70mv." This answer lists all the difference-makers for "action potential being transmitted in the direction of the axon terminal" vs. "action potential being transmitted backwards".

We can apply this framework to behavioral explanations. If we want to explain why some behavior B was observed rather than not, we must compare the mechanism for B with the maximally similar mechanism that does not bring about B. For example, if we want to explain why subjects react to stimuli without any mistakes rather than with many mistakes, we have to compare the mechanism that is active when subjects react to stimuli without mistakes with the maximally similar mechanism that is active when subjects make many mistakes. The differences between these mechanisms will be the difference-makers for B (reaction without mistakes) vs. not-B (reactions with many mistakes).

The crucial idea is that an unconscious phenomenon is to be categorized as "mental" if the difference-makers for B vs. not-B when B is produced by an unconscious phenomenon are also on the list of differences-makers for B vs. not-B when B is produced by a conscious phenomenon. Together with the procedure for finding difference-makers for B vs. not-B, this gives us the following recipe for deciding whether an unconscious phenomenon is mental, or not (B* of type B is the behavior produced by an *unconscious* phenomenon; B** of type B is the behavior produced by a *conscious* phenomenon):

(1) Finding Difference-Makers for B* of Type B vs. not-B: Compare the mechanism underlying B*—call it "M_unconscious"—with the maximally similar mechanism that does not produce any behavior of type B, and list all the differences (*list of difference-makers 1*). These difference-makers explain why the unconscious phenomenon leads to a behavior of type B rather than not.

(2) Finding Difference-Makers for B** of type B vs. not-B: Compare the mechanism underlying B**—call it "M_conscious"—with the maximally similar mechanism that does not produce B, and list all the differences (*list of difference makers 2*). These difference-makers explain why the conscious phenomenon leads to a behavior of type B rather than not.

(3) Same difference-makers? Compare the *list of difference-makers 1* with *list of difference-makers 2*. If these lists overlap, the mechanisms share explanatorily relevant components, and, thus, the conscious and the unconscious phenomenon explain behavior B in a relevantly similar way. Thus, the phenomenon that M_unconscious is responsible for is to be categorized as a mental phenomenon.

To clarify the proposal, let us look at an example. Conscious vision as well as unconscious vision are supposed to explain why subjects react to stimuli with an accuracy greater than chance. Unconscious vision is a mental phenomenon (and thus deserves the label "vision") if the mechanism underlying unconscious vision (partly) consists of those components (or a subset thereof) that are the difference-makers for "reaction with accuracy greater than chance" vs. "chancy reaction" in the case of conscious vision. In order to test, whether this is the case, we have to do different types of experiments and determine the mechanisms that are active in these experimental setups.

Assume we investigate putative unconscious vision with help of a binocular rivalry task as described in Section 2.1. In order to determine whether the phenomenon that is responsible for

the observed behavior really is *vision*, i.e., a mental phenomenon, we have to do the following (where the relevant behavior B is a shift in covert spatial attention):

(i) Finding Difference-Makers for B* of Type B vs. not-B: Compare M_unconsicous underlying the covert attention-shifts triggered by a binocular rivalry task using Mondrians and pictures of nudes that are not consciously perceived (B* of type B) to the maximally similar mechanism active in the same binocular rivalry task that does not lead to B. → *list of difference-makers 1*

(ii) Finding Difference-Makers for B** of type B vs. not-B: Compare M_consicous underlying covert attention-shifts when the stimuli are consciously perceived (B** of type B) with a maximally similar mechanism active in a similar context with consciously perceived stimuli that does not lead to B. → *list of difference-makers 2*

(iii) Same difference-makers? Compare *list of difference-makers 1* and *list of difference-makers 2*: If the components of M_unconscious that make the difference between B* vs. not-B are also components of M_conscious that make the difference between B** vs. not-B, the mechanisms are relevantly similar, i.e., the explain B vs. not-B in a relevantly similar way, and the unconscious phenomenon associated with M_unconscious is to be categorized as "mental".

The condition described by (iii) seems to be satisfied in the case of unconscious vision as investigated in binocular rivalry tasks. Empirical studies suggest that the difference-makers for B (attention-shifts) vs. not-B (no attention shifts) may be early visual cortex activation (Klink and Roelfsema 2016). These difference-makers are involved in M_unconsicous as well as M_consicous. Furthermore, in a recent study, Luo et. al (2021) found that consciously perceived and suppressed stimuli both lead to alpha activation (only that the activation induced by consciously perceived stimuli is stronger). These alpha rhythms that can last for as long as one second are directly related to visual stimulation and thus, would not occur if there was no vision

at all. Furthermore, the alpha activation induced by consciously visible and suppressed stimuli in both cases propagates as a travelling wave from occipital to frontal regions, and they do so with the same strength (Luo, VanRullen, and Alamia 2021, 6). The authors speculate that the main difference is that the alpha activation induced by consciously visible stimuli propagates further in the visual hierarchy, whereas activation generated by the suppressed stimuli disappear at an earlier stage of visual processing (ibid.). However, these higher levels of the visual hierarchy seem to be correlated with conscious higher-level processes rather than to conscious perception per se, as suggested by several studies on vision (such as Kouider et. al (2007); see below). Thus, the mechanisms underlying conscious and unconscious vision as measured in binocular rival tasks seem to share components that are explanatorily relevant for the attention shift. Thus, unconscious vision should be categorized as a mental phenomenon.

The same would have to be shown for (putatively) unconscious vision observed in other experimental setups, such as priming studies. Indeed, there is evidence that the mechanisms underlying subliminal priming and the mechanisms underlying supraliminal priming overlap in the relevant way as well (Kouider et al. 2007). Note that studies investigating (putatively) unconscious vision usually aim at finding the neural correlates of consciousness. Therefore, they aim at describing the *differences* between the mechanism underling the unconscious process and the mechanism underlying the conscious one. Thus, the similarities between the mechanisms are usually not explicitly mentioned. However, the comparison of these two mechanisms is usually framed in a way that suggests that the two mechanisms are taken to be more or less the same minus whatever consciousness adds. For example, in a study comparing brain activity evoked by subliminally and supraliminally presented stimuli, Kouider et al. (2007) conclude that their findings

> fit with the suggestion that subliminal activation remains confined to a few specialized processors, whereas stimuli that pass the visibility threshold become available to a much

> broader range of processors including more anterior cortical sites. (Kouider et al. 2007, 2026)

They also observe that the main difference between the mechanisms triggered by subliminal and supraliminal primes are the activation of a restricted set of occipito-temporal areas, without extension into higher associative regions and suggest that the many additional regions are linked to higher-level cognitive processes such as recognition, report, evaluation, or memory storage rather than the conscious vision as such (Kouider et al. 2007, 2026). Similarly, Railo et. al. (2012) provide evidence that "unconscious perception of color (…) is mediated by early cortical activation in the same time window as conscious vision" (Railo et al. 2012, 828). Thus, priming studies suggest that the mechanisms underlying conscious and unconscious vision are relevantly similar with respect to explaining attention shifts and target detection.

What about implicit attitudes? On the assumption that implicit attitudes are unconscious (see Section 2.2): Are we justified in treating the empirical evidence as evidence for an unconscious mind? To answer this question, we first must determine the relevant mechanisms for the comparison. Assume we use the CV study as presented in Section 2.2 to investigate attitudes. The behavior type B that we want to investigate is (more or less) stable discrimination behavior – in the present study, the likelihood to choose a male candidate over a female candidate despite equal qualifications.

a) Finding Difference-Makers for B* of Type B vs. not-B: Compare M_unconscious underlying the discriminatory behavior in the CV study (B* of type B) with the maximally similar mechanism that does not produce any behavior of type B. → *list of difference-makers 1*)
b) Finding Difference-Makers for B** of type B vs. not-B: Compare M_conscious underlying discriminatory behavior induced by explicit prejudice (B** of type B) in a

similar CV study with the maximally similar mechanism that does not produce B. → *list of difference makers 2*

c) Same difference-makers? Compare the *list of difference-makers 1* with *list of difference-makers 2*. If the lists overlap, the mechanisms share explanatorily relevant components, and, thus, the conscious and the unconscious phenomenon explain behavior B in a relevantly similar way. The phenomenon that M_unconscious is responsible for is to be categorized as a mental phenomenon.

Empirical findings show that the relevant component in M_unconscious that explains why the mechanisms leads to discriminatory behavior rather than not is amygdala activation (Stanley, Phelps, and Banaji 2008). The amygdala is also active in M_conscious and this seems to explain discriminatory behavior as well. (The difference between B* and B** may be due to the fact that in M_unconscious ACC and dPFC are active but not in M_conscious. ACC and dPFC both seem to be involved attenuating amygdala activation (Stanley, Phelps, and Banaji 2008, 167)). Thus, the condition formulated in c) seems to be satisfied, and implicit attitudes turn out to provide evidence for an unconscious mind (assuming that they are indeed unconscious).

Support for my account also comes from considerations regarding the mechanisms underlying consciousness and unconscious processes in general. For example, Lau and Rosenthal (2011) present empirical evidence for higher-order thought theories. They argue that "recent computational modeling work (…) suggests that unconscious and conscious processes probably form a serial hierarchy rather than two parallel channels (…)" (Lau and Rosenthal 2011, 11). All types of models of consciousness that assume this kind of non-parallel approach to conscious vs. unconscious processes in principle provide evidence for my account (if the processing steps that are shared by conscious and unconscious processes indeed are the difference-makers for B vs. not-B).

Is there a negative example, i.e., one where a putative unconscious mental phenomenon turns out to be non-mental? This will be the case for all phenomena explained by dual process models that assume that a given mental activity is realized by two wholly distinct types of mechanisms, depending on whether they are conscious or not. For example, the different memory systems of procedural (unconscious) and declarative (conscious) memory seem to provide a negative example. Both memory systems rely on wholly distinct neural networks (Bayley and Squire 2007; Buffalo and Squire 2015). While declarative memories are encoded in the hippocampus, entorhinal cortex and perirhinal cortex and are consolidated and stored in the temporal cortex and elsewhere, procedural memories are encoded and stored by the cerebellum, putamen, caudate nucleus and the motor cortex. Hence, procedural memory is not a mental phenomenon as it is neither conscious, nor is the mechanism relevantly similar to the mechanism underlying conscious declarative memory. One consequence of this might be that we should not think of skillful behavior (such as playing the piano or riding the bike) that relies on procedural memory as a capacity of the mind. In contrast, *learning* a skill may be as this usually involves consciousness. However, I will leave a defense of these claims for another paper.

4.5 Why mechanisms?

Why should similarity between *mechanisms* be the criterion for unconscious mentality? As argued above, one reason is that conscious and unconscious mental phenomena are postulated as explanatory posits that play the same role in *mechanistic explanation* of certain behaviors. There is a further argument. As argued by Taylor (2020), mental phenomena such as perception are taken to be natural kinds (see also Shepherd and Mylopoulos (2021)). Unconscious phenomena are mental if and only if they belong to a mental natural kind. As Taylor stresses, the most prominent approach to natural kinds is Boyd's homeostatic property cluster view (Boyd 1989): natural kinds are clusters of properties that are held together by an underlying

mechanism. Thus, sharing the same mechanism is a necessary condition for unconscious processes to be of the same natural kind as their conscious counterparts. As different instantiations of the same mechanism bring about the same cluster of properties, my account is in line with Taylor's views.

One potential objection against the claim that my account is in line with the homeostatic property cluster view is that, according to my proposal, the mechanisms underlying conscious phenomena and those underlying unconscious mental phenomena are not the same – they only share the difference-makers for B vs. not-B. This, however, is not sufficient to be of the same natural kind. Two thoughts on this objection: first, my account could be used to improve the homeostatic property cluster view. The view is confronted with the problem that it is unclear how similar mechanisms need to be in order to count as the same mechanism and how similarity is to be understood (see, e.g., Craver (2009)). Similarity in the homeostatic property cluster view could be spelled out in terms of explanatory relevance along the lines of my account. Second, the connection between natural kinds and "similar but not the same" mechanisms (or mechanistic models) is highlighted in Rob Rupert's (2019) "tweak and extend" account of natural kinds. According to this account, the instances of a natural kind bear a family resemblance to each other that is "determined by the causal-explanatory roles of the components" of these instances (Rupert 2019, 37). If two phenomena cannot be modelled by the same model even after tweaking or extending the model, then it turns out that the two phenomena do not belong to the same natural kind. Developing these ideas further will require more space – and therefore will be left for future research.

5 Conclusion

I provided an answer to the *conceptual unconscious mind worry*, i.e., the question of what justifies calling unconscious phenomena "mental". This account has several implications for the *methodological unconscious mind worry,* i.e., the question of how to empirically establish

whether a behavior is due to an unconscious mental phenomenon or not. First, the mechanisms that we must compare are plausibly neurobiological ones. Neuroscience will, thus, be important for settling the question of whether a given behavior is due to an unconscious mental phenomenon, or not. Note that this does not imply that mechanisms need to be described in neurobiological terms. As argued by many new mechanists, mechanisms can also be described in, for example, computational or mathematical terms (Craver and Kaplan 2020). This brings with it the crucial insight that similarities between conscious and unconscious mechanisms may show up on, say, the computational level but may be invisible at the neurobiological level. Hence, computational neuroscience and other sciences that make use of non-biological descriptions of the brain will be relevant, too. Note further that it may turn out that the mechanisms for some mental phenomena exceed the brain (Clark and Chalmers 1998). This possibility has been discussed in the context of mechanistic explanation (Kaplan 2012; Krickel 2020). The account suggested here is compatible with this view. If the extended mind thesis turns out to be true (for at least some mental mechanisms), other sciences may be important for addressing the methodological unconscious mind worry as well.

Second, to show that a behavior is due to an unconscious mental process, a lot of experimentation and knowledge is necessary. Based on the proposed account, at least in principle, four mechanisms (one leading to B*, one leading to B**, and the two maximally similar ones that do not lead to B) must be considered. Note that we do not have to have complete knowledge about the mechanisms underlying the relevant phenomena. We only must establish that there is *one* mechanistic component that is shared by M_unconscious and M_conscious that is not present in the respective mechanisms not producing B.

Finally, whether an unconscious phenomenon is mental or not turns out to be an empirical question rather than one that can be answered on a priori grounds. This highlights that answering remaining questions regarding the methodological unconscious mind worry is an

important task. Remaining questions are: Which experimental designs are most adequate to investigate a given mental phenomenon? What if different experimental designs investigating the same mental phenomenon deliver different results regarding the underlying mechanisms? Which level of description (neurobiological, computational, mathematical, …) is the correct one to detect similarities and differences between mechanisms? The present proposal can serve as a starting point for thinking about these questions.

References

Axt, Jordan R. 2018. "The Best Way to Measure Explicit Racial Attitudes Is to Ask About Them." *Social Psychological and Personality Science* 9 (8): 896–906. https://doi.org/10.1177/1948550617728995.

Banaji, Mahzarin R., R. Bhaskar, and Michael Brownstein. 2015. "When Bias Is Implicit, How Might We Think about Repairing Harm?" *Current Opinion in Psychology* 6 (December): 183–88.

Bargh, John A., and Ezequiel Morsella. 2008. "The Unconscious Mind." *Perspectives on Psychological Science* 3 (1): 73–79. https://doi.org/10.1111/j.1745-6916.2008.00064.x.

Bayley, Peter J., and Larry R. Squire. 2007. "The Neuroanatomy and Neuropsychology of Declarative and Nondeclarative Memory." In , 1–18. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-45702-2_1.

Bechtel, William, and Adele Abrahamsen. 2005. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421–41. https://doi.org/10.1016/j.shpsc.2005.03.010.

Berger, Jacob. 2014. "Mental States, Conscious and Nonconscious." *Philosophy Compass* 9 (6): 392–401. https://doi.org/10.1111/phc3.12140.

Berger, Jacob, and Bence Nanay. 2016. "Relationalism and Unconscious Perception." *Analysis* 76 (4): 426–33.

Boyd, Richard. 1989. "What Realism Implies and What It Does Not." *Dialectica* 43 (1/2): 5–29. http://www.jstor.org/stable/42970608.

Brownstein, Michael, Alex Madva, and Bertram Gawronski. 2019. "What Do Implicit Measures Measure?" *Wiley Interdisciplinary Reviews: Cognitive Science* 10 (5): 1–13. https://doi.org/10.1002/wcs.1501.

Buffalo, Elizabeth A., and Larry R. Squire. 2015. "Declarative Memory, Neural Basis Of." In *International Encyclopedia of the Social & Behavioral Sciences*, 923–26. Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.51005-8.

Burge, Tyler. 2010. *The Origins of Objectivity*. Oxford: Oxford University Press.

Carruthers, Peter. 2015. *The Centered Mind*. Oxford: Oxford University Press.

Carter, J. Adam, and Robert D. Rupert. 2020. "Epistemic Value in the Subpersonal Vale." *Synthese*, no. May 2019. https://doi.org/10.1007/s11229-020-02631-1.

Clark, Andy, and David J Chalmers. 1998. "The Extended Mind." *Analysis* 58 (1): 7–19.

Craver, Carl F. 2007. *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press.

———. 2009. "Mechanisms and Natural Kinds." *Philosophical Psychology* 22 (5): 575–94. https://doi.org/10.1080/09515080903238930.

Craver, Carl F., and David M. Kaplan. 2020. "Are More Details Better? On the Norms of Completeness for Mechanistic Explanations." *The British Journal for the Philosophy of Science* 71 (1): 287–319. https://doi.org/10.1093/bjps/axy015.

Degner, Juliane, Dirk Wentura, Burkhard Gniewosz, and Peter Noack. 2007. "Hostility-Related Prejudice against Turks in Adolescents: Masked Affective Priming Allows for a Differentiation of Automatic Prejudice." *Basic and Applied Social Psychology* 29 (3):

245–56. https://doi.org/10.1080/01973530701503150.

Dennett, Daniel C. 2007. “Philosophy as Naive Anthropology: Comment on Bennett and Hacker.”

Dijksterhuis, Ap, and Loran F. Nordgren. 2006. “A Theory of Unconscious Thought.” *Perspectives on Psychological Science* 1 (2): 95–109. https://doi.org/10.1111/j.1745-6916.2006.00007.x.

Drayson, Zoe. 2012. “The Uses and Abuses of the Personal/Subpersonal Distinction.” *Philosophical Perspectives* 26 (1): 1–18. https://doi.org/10.1111/phpe.12014.

Feest, Uljana. 2020. “Construct Validity in Psychological Tests – the Case of Implicit Social Cognition.” *European Journal for Philosophy of Science* 10 (1): 4. https://doi.org/10.1007/s13194-019-0270-8.

Figdor, Carrie. 2018. *Pieces of Mind*. *Pieces of Mind: The Proper Domain of Psychological Predicates*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198809524.001.0001.

Gawronski, Bertram, Wilhelm Hofmann, and Christopher J. Wilbur. 2006. “Are ‘Implicit’ Attitudes Unconscious?” *Consciousness and Cognition* 15 (3): 485–99. https://doi.org/10.1016/j.concog.2005.11.007.

Glennan, Stuart. 2017. *The New Mechanical Philosophy*. Oxford University Press.

Greenwald, Anthony G, Debbie E Mcghee, and Jordan L K Schwartz. 1998. “Measuring Individual Differences in Implicit Cognition: The Implicit Association Test.” *Journal of Personality and Soclal Psychology* 74 (6): 1464–80. https://doi.org/10.1037/0022-3514.74.6.1464.

Hahn, Adam, Charles M Judd, Holen K Hirsh, and Irene V Blair. 2014. “Awareness of Implicit Attitudes.” *Journal of Experimental Psychology: General* 143 (3): 1369–92. https://doi.org/10.1037/a0035028.

Hassin, Ran R. 2013. "Yes It Can: On the Functional Abilities of the Human Unconscious." *Perspectives on Psychological Science* 8 (2): 195–207. https://doi.org/10.1177/1745691612460684.

Illari, Phyllis McKay, and Jon Williamson. 2012. "What Is a Mechanism? Thinking about Mechanisms across the Sciences." *European Journal for Philosophy of Science* 2 (1): 119–35. https://doi.org/10.1007/s13194-011-0038-2.

Jiang, Y., P. Costello, F. Fang, M. Huang, and S. He. 2006. "A Gender- and Sexual Orientation-Dependent Spatial Attentional Effect of Invisible Images." *Proceedings of the National Academy of Sciences* 103 (45): 17048–52. https://doi.org/10.1073/pnas.0605678103.

Johnson, Gabbrielle M. 2019. "The Structure of Bias." *Mind* 129 (516): 1193–1236. https://doi.org/10.1093/mind/fzaa011.

Kaplan, David M. 2012. "How to Demarcate the Boundaries of Cognition." *Biology and Philosophy* 27 (4): 545–70. https://doi.org/10.1007/s10539-012-9308-4.

Kästner, Lena, and Lise Marie Andersen. 2018. "Intervening into Mechanisms: Prospects and Challenges." *Philosophy Compass* 13 (11): e12546. https://doi.org/10.1111/phc3.12546.

Klink, P. Christiaan, and Pieter R Roelfsema. 2016. "Binocular Rivalry Outside the Scope of Awareness." *Proceedings of the National Academy of Sciences* 113 (30): 8352–54. https://doi.org/10.1073/pnas.1609314113.

Kohár, Matej, and Beate Krickel. 2021. "Compare and Contrast: How to Assess the Completeness of Mechanistic Explanation." In *Neural Mechanisms - New Challenges in the Philosophy of Neuroscience*, edited by Fabrizio Calzavarini and Marco Viola, 395–424. Springer. https://doi.org/10.1007/978-3-030-54092-0_17.

Kouider, Sid, Stanislas Dehaene, Antoinette Jobert, and Denis Le Bihan. 2007. "Cerebral Bases of Subliminal and Supraliminal Priming during Reading." *Cerebral Cortex* 17 (9):

2019–29. https://doi.org/10.1093/cercor/bhl110.

Krickel, Beate. 2020. “Extended Cognition, the New Mechanists’ Mutual Manipulability Criterion, and the Challenge of Trivial Extendedness.” *Mind & Language* 35 (4): 539–61. https://doi.org/10.1111/mila.12262.

Lau, Hakwan, and David Rosenthal. 2011. “Empirical Support for Higher-Order Theories of Conscious Awareness.” *Trends in Cognitive Sciences* 15 (8): 365–73. https://doi.org/10.1016/j.tics.2011.05.009.

Luo, Canhuang, Rufin VanRullen, and Andrea Alamia. 2021. “Conscious Perception and Perceptual Echoes: A Binocular Rivalry Study.” *Neuroscience of Consciousness* 2021 (1): 1–8. https://doi.org/10.1093/nc/niab007.

Mandelbaum, Eric. 2016. “Attitude, Inference, Association: On the Propositional Structure of Implicit Bias.” *Nous* 50 (3): 629–58. https://doi.org/10.1111/nous.12089.

Morgan, Alex, and Gualtiero Piccinini. 2018. “Towards a Cognitive Neuroscience of Intentionality.” *Minds and Machines* 28 (1): 119–39. https://doi.org/10.1007/s11023-017-9437-2.

Newell, Ben R, and David R Shanks. 2014. “Unconscious Influences on Decision Making: A Critical Review.” *Behavioral and Brain Sciences* 38 (01): 1–19. https://doi.org/10.1017/S0140525X12003214.

Persuh, Marjan. 2018. “The Fata Morgana of Unconscious Perception.” *Frontiers in Human Neuroscience* 12 (April): 1–5. https://doi.org/10.3389/fnhum.2018.00120.

Peters, Megan A. K., Robert W. Kentridge, Ian Phillips, and Ned Block. 2017. “Does Unconscious Perception Really Exist? Continuing the ASSC20 Debate.” *Neuroscience of Consciousness* 3 (1): 1–11. https://doi.org/10.1093/nc/nix015.

Phillips, Ian. 2018. “Unconscious Perception Reconsidered.” *Analytic Philosophy* 59 (4): 471–514. https://doi.org/10.1111/phib.12135.

Phillips, Ian, and Ned Block. 2016. “Debate on Unconscious Perception.” In *Current Controversies in Philosophy of Perception*, edited by Bence Nanay, 165–92. London: Routledge.

Piccinini, Gualtiero, and Carl F. Craver. 2011. “Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches.” *Synthese* 183 (3): 1–58. https://doi.org/10.1007/s11229-011-9898-4.

Pothos, Emmanuel M. 2007. “Theories of Artificial Grammar Learning.” *Psychological Bulletin* 133 (2): 227–44. https://doi.org/10.1037/0033-2909.133.2.227.

Prinz, Jesse. 2015. “Unconscious Perception.” In *The Oxford Handbook of Philosophy of Perception*, edited by Mohan Matthen, 371–91. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199600472.013.033.

Quilty-Dunn, Jake, and Eric Mandelbaum. 2018. “Against Dispositionalism: Belief in Cognitive Science.” *Philosophical Studies* 175 (9): 2353–72. https://doi.org/10.1007/s11098-017-0962-x.

Railo, Henry, Niina Salminen-Vaparanta, Linda Henriksson, Antti Revonsuo, and Mika Koivisto. 2012. “Unconscious and Conscious Processing of Color Rely on Activity in Early Visual Cortex: A TMS Study.” *Journal of Cognitive Neuroscience* 24 (4): 819–29. https://doi.org/10.1162/jocn_a_00172.

Rupert, Robert D. 2018. “The Self in the Age of Cognitive Science: Decoupling the Self from the Personal Level.” *Philosophic Exchange* 47 (1): 1–36.

———. 2019. “Group Minds and Natural Kinds.” *Avant* 10 (3): 1–28. https://doi.org/10.26913/avant.2019.03.08.

Saul, Jennifer. 2013. “Implicit Bias, Stereotype Threat, and Women in Philosophy.” In *Women in Philosophy*, 39–60. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199325603.003.0003.

Searle, John R. 1992. *The Rediscovery Of The Mind*. MIT Press. https://doi.org/10.1016/0004-3702(95)90037-3.

Shepherd, Joshua, and Myrto Mylopoulos. 2021. "Unconscious Perception and Central Coordinating Agency." *Philosophical Studies*. https://doi.org/10.1007/s11098-021-01629-w.

Smith, Ryan, and Richard D. Lane. 2016. "Unconscious Emotion: A Cognitive Neuroscientific Perspective." *Neuroscience & Biobehavioral Reviews* 69 (October): 216–38. https://doi.org/10.1016/j.neubiorev.2016.08.013.

Stanley, Damian, Elizabeth Phelps, and Mahzarin Banaji. 2008. "The Neural Basis of Implicit Attitudes." *Current Directions in Psychological Science* 17 (2): 164–70. https://doi.org/10.1111/j.1467-8721.2008.00568.x.

Taylor, Henry. 2020. "Fuzziness in the Mind: Can Perception Be Unconscious?" *Philosophy and Phenomenological Research* 101 (2): 383–98. https://doi.org/10.1111/phpr.12592.

Thomson, Eric, and Gualtiero Piccinini. 2018. "Neural Representations Observed." *Minds and Machines* 28 (1): 191–235. https://doi.org/10.1007/s11023-018-9459-4.

Uhlmann, Ericluis, and Geoffrey L. Cohen. 2005. "Constructed Criteria: Redefining Merit to Justify Discrimination." *Psychological Science* 16 (6): 474–80. https://doi.org/10.1111/j.0956-7976.2005.01559.x.

Weiskrantz, Lawrence. 1995. "Blindsight: Not an Island Unto Itself." *Current Directions in Psychological Science* 4 (5): 146–51.

Yang, Eunice, David H Zald, and Randolph Blake. 2007. "Fearful Expressions Gain Preferential Access to Awareness during Continuous Flash Suppression." *Emotion* 7 (4): 882–86. https://doi.org/10.1037/1528-3542.7.4.882.