Unconscious Perception and Unconscious Bias: Parallel Debates About Unconscious Content

Gabbrielle M Johnson

Abstract

The possibilities of unconscious perception and unconscious bias prompt parallel debates about unconscious mental content. In this article, I argue that claims within these debates alleging the existence of unconscious content are made fraught by ambiguity and confusion with respect to the two central concepts they involve: consciousness and content. Borrowing conceptual resources from the debate about unconscious perception, I distill the two conceptual puzzles concerning each of these notions and establish philosophical strategies for their resolution. I then argue that empirical evidence for unconscious bias falls victim to these same puzzles, but that progress can be made by adopting similar philosophical strategies. Throughout, I highlight paths forward in both debates, illustrate how they serve as fruitful domains in which to study the relationship between philosophy and empirical science, and use their combined study to further understanding of a general theory of unconscious content.

1 Introduction

Two parallel debates about unconscious mental content serve as fruitful domains in which to study the relationship between philosophy and empirical science. The first debate, about unconscious perception, highlights conceptual confusion regarding its two central notions of *consciousness* and *perception.* Progress in understanding the nature of unconscious perception is due to a mutual exchange between philosophers and psychologists working in tandem to clarify concepts, test those concepts using empirical methods, and use the combined conceptual-empirical methods to make progress on a general theory of unconscious content. A second debate, emerging more recently, concerns the existence of unconscious social bias. Though still in early stages, comparatively less attention has been paid to the possible exchange between philosophy and social psychology. But, as I'll argue, here too empirical science and philosophy have much to gain from one another.

In this article, I zoom in on these two contemporary debates concerning unconscious content in order to illustrate the benefits of exchange between philosophy and empirical science. I begin, in section two, with some conceptual groundwork, introducing two philosophical marks of the mental that will be central to each debate to follow—consciousness and representational content—and associated problems that accompany empirical study of both. Then, in section three, I investigate the debate around unconscious perception, demonstrating how it has clarified the conceptual problems facing any debate about unconscious content and how it has attempted to overcome these problems. In section four, I investigate the debate of unconscious social bias, demonstrating how it is subject to these same conceptual problems and arguing that progress on them can be made by adopting lessons learned from the debate about unconscious perception.

2 Consciousness and Content

Historically, philosophers have found it helpful to distinguish between two so-called "marks of the mental": *consciousness* and *intentional content*. Consciousness is paradoxically the most familiar and the most mysterious mark of the mental. Most famously, the notion of consciousness is thought to be captured by "the feeling of what it's like." If some creature is conscious, then there is a feeling of what it is like to be that creature. If some state of the creature is conscious, then there is something it is like for the creature to be in that state. That is, when in this state, the creature has some unique, subjective perspective on how the world is presented to them. Consider the feeling of what it's like to stub your toe, to wake up to the rich smell of coffee, to stare intently at a perfectly ripe tomato, to fall in love. Each of these experiences is accompanied by a rich host of phenomenal qualities (or 'qualia') that are presented to you and only you through your subjective experience.[1] In a nutshell, consciousness corresponds roughly to the inward-directed, subjective aspects of mental life: how things feel *for you*.[2]

[1] More precisely, we might follow Kriegel 2009 in distinguishing between the "subjective character" of there being something it's like to be in the state and the "qualitative character" of what it's like to be in the state.

[2] As much as possible, I avoid making distinctions between different kinds of consciousness (access, phenomenal, introspective, etc.). See Block 1995 and Berger 2022 for further, helpful distinctions. I likewise avoid spelling out theories of what makes something consciously accessible. This unfortunately includes

Intentional content, on the other hand, corresponds roughly to the mind's ability to reach out into the world, bringing outward objects into the mind to think about (and, perhaps, present through conscious experience). Historically, philosophers have associated intentional content with *aboutness*. My belief that Barack Obama attended Harvard Law School has this quality: it is about things out there in the world, most centrally a person named 'Barack Obama', a law school named 'Harvard Law School', and the relation of *attending*. We might say that the representational content of my belief is, again roughly, the proposition that "Barack Obama attended Harvard Law School". Crucially, the way the world is represented might not be how the world actually is—perhaps Barack Obama attended Yale. In these cases, the representational content is false or inaccurate.[3]

The view of the mind we're left with is one where the mind is conceptually split into two capacities: consciousness (i.e., subjective experience) and cognition (i.e., mental calculation involving representational contents). Purported cases of unconscious content to be discussed will involve states that allegedly have just one of these marks: content. That is, they lack consciousness but have representational content. This involves two separate arguments: first, that they're unconscious; and second that they're representational. Each debate to follow presents challenges to each step.

In what follows, I walk through each of these debates (about consciousness and content) for each case study (about perception and bias) totaling four dialectical touchstones. This will highlight the extensions from one to the other: arguments surrounding consciousness of perception will be similar to, and help to inform, arguments surrounding consciousness for bias, likewise for arguments surrounding content for both.

Regarding consciousness, the debates in both domains will boil down to the difficulty in establishing that some state is truly unconscious. This is partly due to there being no universally agreed-upon empirical methods for studying consciousness.[4] In most cases,

interesting theories specifically with respect to the debates at hand, e.g., HOT theories (Berger 2020, Rosenthal 2005), inferential/interpretive awareness theories (Carruthers 2017), and attention, categorization, and control of action theories (Krickel 2018), all of which are developed in the context of the debate about unconscious social bias. Questions about what makes some state consciously accessible are different from the question of how we might empirically test for consciousness, the latter of which will be a central theme of the paper.

[3] Here I am presenting a view that roughly accords with a representational theory of mind. Not all philosophers subscribe to such a view of the mind, but the framework is nearly ubiquitous in empirical psychology and cognitive science. So adopting it here has the additional advantage of facilitating communication between the two domains, which is a primary goal of the paper.

[4] Some question whether there can ever be a satisfying scientific explanation of consciousness. See Irvine 2013 for a comprehensive discussion. I avoid such extreme skepticism here.

empirical methods purporting to study consciousness rely on subjective reports. However, notoriously, subjective reports are prone to response bias. Thus, in any particular case, critics of unconscious content will argue that purported cases are not truly unconscious, but rather cases where subjective report is in some sense obscuring the conscious accessibility of the state. I'll call the broad collection of these sorts of issues "the Conscious Criterion Problem":[5]

> **The Conscious Criterion Problem:** Some purported unconscious representational content might in fact not be unconscious (because it is a conscious state registering below a subjective response criterion).

Likewise, in the absence of consciousness, establishing that a state has representational content is not a straightforward matter. This is because we standardly distinguish between a robust sort of intentionality, exhibited by mental states that are attributable to individuals, and weaker statistical notions, exhibited when parts of the natural world carry information about other parts of the natural world. The natural world finds plenty of ways for some things to carry information about other things. This can be as mundane as the angle of a rock's shadow carrying information about the time of day. If this is all there were to intentional content, then it might seem that the mind's capacity for *aboutness* is not all that unique, i.e., is no longer a mark of the mental. Thus, in any particular case, critics of unconscious content will argue that purported cases are not truly representational contents attributable as mental states to the individuals that harbor them, but rather cases where there is some low-level, non-representational information processing occurring. In essence, they're arguing that such processes' lacking consciousness is theoretically uninteresting, as it's tantamount to a rock's shadow not being accompanied by subjective experience. I'll call the broad collection of these sorts of issues, "the Content Attribution Problem":

> **The Content Attribution Problem:** Some purported unconscious representational content might in fact not be representational content (because it does not differ significantly from low-level, non-representational information processing).

[5] My Conscious Criterion Problem and Content Attribution Problem are modeled roughly on Ian Phillips's Criterion and Attribution Problems to be discussed below. Phillips's Attribution problem focuses more on so-called "attributability to the individual" than mine does. My reasons for this departure will become clear as the debates unfold.

I now turn to the two parallel debates on unconscious content that center around these two concepts and their associated problems.

3 Is unconscious perception unconscious content?

The first purported case of an unconscious representational content is unconscious perception. Following tradition in philosophy and psychology, the focus here will be almost exclusively on visual perception. Exploration of this case will serve as an exemplary model of how philosophers and scientists can mutually inform our understanding of unconscious content.

3.1 What is perception?

For the question of whether perception can be unconscious to make sense, we need some characterization of perception that does not presuppose conscious accessibility. If we were to define visual perception as conscious visual awareness, then since there can be no unconscious conscious visual awareness, there could be no unconscious perception. Thus, we need some characterization that is independent of consciousness.

One such independent characterization utilized in debates about unconscious perception originates in the work of Tyler Burge. Burge (2010, 2022) characterizes perception as **objective sensory representation, paradigmatically by the individual**. In what follows, I summarize each element of this characterization using a concrete (though idealized) example.

Suppose I have an accurate visual perception of a red berry under blue light. The basic mechanics of how this visual perception comes to fruition begins with the distal cause of my perception (the red berry) and culminates in the perceptual state itself (with the content "red berry"). In between, there is a complex chain of non-perceptual, causal processes linking the two. The chain begins with blue light streaming down and coming into contact with the red surface of the berry (what we'll call *the distal cause*). At that point, a combination of blue and red light bounces off the berry, resulting in purple. To keep track, we'll call the blue light *the illuminance*, the red *the reflectance*, and the combination of purple *the luminance*. The purple luminance is what the eyes register initially, on a two-dimensional (2-D) array constituting the retina. We regard it as the *proximal stimulus* because it is closer than the *distal* cause. When the luminance makes contact with the array, the sensory registrations responsible for activating when hit with the appropriate kind of stimulus (in this case, the purple light) activate. Thus, the sensory array registers information about the luminance. But notice the puzzle here: what we've registered is the purple light of the proximal stimulus, but what we're really interested in is the color of the distal berry itself, which is red. So, the visual system's primary function is to get us from proximal to distal. It does this by processing the initial sensory information through a series of rule-governed calculations, one of which will subtract surrounding light for the purpose of getting at the true color of the berry. Thus, this calculation takes as an input the luminance (purple) and subtracts the illuminance (blue) so as to produce an output of the reflectance (red). The whole process is a paradigm instance of cognition (mental calculation involving representational contents):

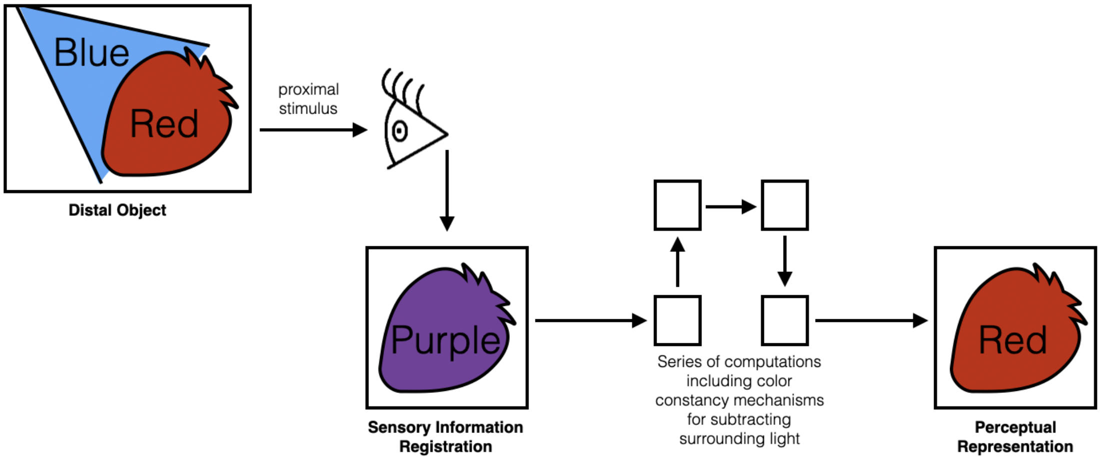

Blue
Red
Distal Object
proximal stimulus
Purple
Sensory Information Registration
Series of computations including color constancy mechanisms for subtracting surrounding light
Red
Perceptual Representation

Figure One: An idealized model of seeing a red berry under blue light.

Crucially, the same type of sensory information registration could be the result of many possible distal objects. For example, the same registration of purple could be the result of a blue berry under red light or a purple berry under purple light (and there are indefinitely many other combinations as well). So there are many other cues and rule-governed processes that the visual system uses to make a best guess as to how to perform the calculation correctly. Discovering which cues are important and what rules govern the processes is the primary aim of vision science. This is a difficult task and made possible only by exploring a myriad of cases where things go awry. While the visual system is astonishingly good at getting things right, exploring cases where it gets things wrong gives psychologists hints about what sorts of cues and assumptions it relies on.

In the berry example, a nonstandard visual scene, such as an image cleverly designed to seem like uniform blue illumination, but that in fact has an illumination gap just where the berry is, will result in a perceptual illusion. In this case, the visual system's assumption that illuminance is uniform will cause it to subtract blue light from the berry. But since there is no illumination on the berry, subtracting blue will lead it astray. Subtracting in this context makes the distal object appear redder than it actually is, resulting in visual illusions.[6] Working backwards, scientists observe that we're systematically prone to illusions in these sorts of contexts, and so infer to internal workings of the visual system, i.e., the assumption that illumination is uniform and the process that subtracts it. Thus, knowing the cases of when the visual system gets things right (produces accurate representations) and when it gets things wrong (produces inaccurate representations) is crucial to vision science. In this way, a notion of intentional or representational content is baked into the very heart of contemporary cognitive psychology.

With this example in hand, we can inspect each of the four aspects of Burge's characterization: that perception is **objective**, **sensory**, **representational** that is **paradigmatically by the individual**. Importantly, this characterization is not meant to give collective necessary or sufficient conditions for some psychological state's being perception.[7] The notions of *objective* and *representational* are most central to Burge's understanding of perception, and are most important for our purposes of extending lessons from the debate about perception to the debate about bias.

Let's start with the claim that perceptual states are **sensory**. This means that perceptual states, unlike other higher-level cognitive states, are constitutively tied to sensory modalities. Visual perception is tied to vision, auditory perception to audition, etc. A state's being tied to sensory modalities in this way can be contrasted with higher-level cognitive states that are amodal. Propositional attitudes like belief and desire seem not tied to any specific modality. Some states can be sensory but not be perception. The information registrations at the retina are an example. These states are not perception because they do not rise to the status of objective representation.

[6] You can explore these sorts of illusions for yourself here: https://www.echalk.co.uk/amusements/OpticalIllusions/colourPerception/colourPerception.html.

[7] Burge 2022, 19. There are some psychological states that are objective sensory representations by the individual that are not perception, such as states of perceptual memory and anticipation (these are perceptual, but not perception, according to Burge). However, it will be impossible that some state is not objective or representational, but is still perception.

Which bring us to the two most central notions of the characterization: objectivity and representation. For a state to be **representational** is, for Burge, for veridicality conditions (namely, conditions under which we would say the state is true or accurate) to feature as an essential aspect of its nature. In other words, veridicality conditions are essential to that state's being the state that it is. Contrasts are helpful. In the case of the information registration, that state is characterized in straightforward causal terms: the registrations were directly caused by the light hitting the retina. Of course, we can talk about the state *as if* it had veridicality conditions (we could say that the retina "represents" purple light), but we don't have to. In this case, simpler non-representational, mere causal statements can do the trick. Contrast that with the representation of red at the end of the process. Here, a direct causal story won't do. This is partly because the red berry is not a direct cause of the perception; rather, it is mediated by a variety of other causal links (including the sensory registration). Likewise, while it might be true in some cases to say that the representation was distally caused by the red reflectance of the berry, it needn't always be caused in this way. As we saw in the case of optical illusions, the very same type of sensory registration, producing the very same type of representational state, could be formed by very different distal causes. Notice that in these cases, e.g., where the reflectance of the berry is purple but we inaccurately form the representation of a red reflectance, we cannot point to a red reflectance as the cause (not even distally). Thus, to characterize the nature of the representational state, we *must* go beyond reference to direct, proximal causes. We *must* make use of veridicality conditions, saying what the state was aiming to pick out (even in the absence of that thing). This is roughly what it means for the state to have veridicality conditions as an aspect of the state's nature, i.e., for the state to be representational.

This discussion also helpfully elucidates what it means for the state to be **objective**. In this case, the representational state is objective insofar as it captures features of the distal object that go beyond idiosyncratic features of the subject's registrational capacities. In other words, the representation can continue to represent the distal berry as red, despite the ever-changing idiosyncratic registrations at the retina. Were we to move the red berry out of blue light and into yellow light, the proximal registrations would change dramatically (shifting from purple to orange). However, we wouldn't thereby see the berry as drastically changing color. Instead, we continue to see the berry as red. Our visual system first subtracts the blue, and then subtracts the yellow, keeping the representation red. This ability to keep distal features constant despite ever-changing proximal stimulations is precisely the function of what are called “perceptual constancy mechanisms”. Because of their ability to facilitate the tracking of distal features, resulting in the formation of objective representational states, the constancy mechanisms posited by visual psychology feature centrally in Burge’s theory of perception. They take us from sensory informational registrations that are non-representational and merely causal to rich representational states that are of (or *as of*, in the case of misrepresentation) constant, distal features of the environment. They are, according to Burge, the origins of objectivity in the mind. They are what allow representational mind to begin.

What of the last element of the characterization, that perception is **paradigmatically by the individual**? One might think this is the most central aspect of the characterization given how prominently it figures in the debates to follow. However, I’ll argue that this emphasis is misplaced, and that rather we should maintain a focus on the two most central notions of the characterization: objectivity and representation. Still, it will help to give a brief gloss of what is meant by the idea.

Formally, a state’s being *by the individual* means that it is “functionally integrated with the exercises of whole individual functions.”[8] The basic distinction Burge is interested in is between states and processes that it is natural to attribute to the individual as opposed to those it is natural to attribute merely to an individual’s subsystems. Some cases can be illustrated by consideration of simple intuitions. For example, does it make more sense to say that I subtract surrounding light from the information registration or that my visual system does? Does it make more sense to say that I perceive a red berry or that my visual system does? Following these intuitions, perceptual constancies are attributed to the perceptual system whereas the representational states themselves are attributed to the individual.[9] Perception, according to Burge, is sometimes (and paradigmatically) by

[8] Burge 2020, 69, fn. 45.

[9] Burge never gives clear criteria for attribution to the individual. Other considerations beyond intuition are varied and show up primarily in orthogonal discussions regarding behavior and primitive agency (see Burge 2010, 333-341, 369-376 for full discussion). Some considerations (Burge 2010, 331, 333) rest on a state or process issuing from “central [behavioral/coordinating] capacities”, a feature Phillips (Phillips and Block 2017, 169, 174, 181, Phillips 2018, 494, 497-499) in the debate to follow places a lot of weight on, but is ultimately inessential to Burge’s notion. Other considerations of Burge’s involve whether the state involves the whole body of the individual or, more intuitively, just the head (Burge 2010, 332, fn. 60). Some states and processes can be attributed to the individual without being representational or perceptual. For example, a feeling of pain might be entirely sensory (non-representational), but it is felt by the individual. Crucially, some states and processes can be attributable to the individual without being guided by the individual (i.e., under conscious, deliberative control of the individual). Burge (2010, 332) gives the example of a creature writhing in pain as a pure reflex that is nevertheless attributable to the individual. Burge says explicitly that action guided by perception need not be under the control or guidance of the individual, insofar as

the individual, but he allows that some perceptual states could occur that are not by the individual, e.g., some lower-levels of visual processing might be better attributed only to the visual system. We will return to these issues again in the debates to follow.

With this characterization of perception in hand, we can now turn to evidence for the existence of unconscious perception.

3.2 What is unconscious perception?

I begin with two cases frequently offered to support the claim that unconscious visual perception exists—blindsight (or neglect) and continuous flash suppression (involving subliminal priming). I will follow the trajectory of the debate as it has been laid out by Ian Phillips and Ned Block, two empirically-oriented philosophers of mind. Block defends the position that unconscious perception exists and is well-established by scientific findings. Phillips denies the existence of unconscious perception, providing both a systematic critique of the purported evidence as well as philosophical grounds on which to doubt convincing evidence could ever be marshaled.[10]

The first purported cases of unconscious perception, blindsight and neglect, involve partial damage to areas of the brain responsible for visual processing. This results in instances where subjects report having no conscious awareness of at least some part of their visual field, but are nevertheless able to perform behavioral similar to normally-sighted individuals.

Consider a famous case of neglect studied by Marshall and Halligan (1988). The subject P.S. was presented with the following image, aligned in such a way that the left sides of both houses fell into her left visual field (her "blind" side):

"the individual could not monitor or adjust" the action (Burge 2010, 335, fn. 62). In the surrounding context of that footnote, Burge provides an extensive discussion of a grouse's reflexive, instinctual behavior to mate. This behavior is guided purely by visual perception and not the grouse in some voluntary way, but attributable to the grouse nevertheless. Likewise, he gives the example of an individual's ducking reflex being attributable to them, but not under voluntary control (Burge 2010, 335). Burge (2010, 369) is also explicit that not all perception is by the individual. Rather, he says, perception is *fundamentally* by the individual, meaning that all perceptions either are themselves—or contribute to other perceptions that are themselves—attributable to the individual. For more discussion, see Burge 2020, fn. 45.

[10] The debate ranges over a series of articles including, but not limited to, Block 2015, Phillips 2015, Phillips and Block 2017, Phillips 2018, and Phillips 2021.

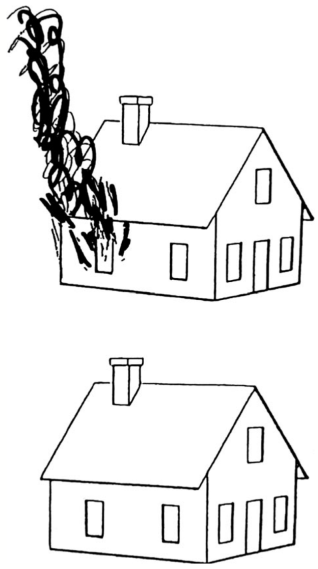


Figure Two: Stimulus Used by Marshall and Halligan (1988) to Demonstrate Neglect.

P.S.'s reports suggested that she lacked conscious access to the left side of both houses. When asked whether the houses are the same or different, she responded that they are the same. When asked if there was anything wrong with either of the houses, she responded no. These responses seem to vindicate claims that she lacked conscious awareness on this side. However, when she was asked which house she would prefer to live in, she chose the house without fire well above chance (on 9 out of 11 trials).[11] This seems to indicate that she represented features of the left side of the visual scene, despite purportedly lacking conscious awareness of it. Thus, it seems a candidate case of perceptual representation without consciousness.

Some, like Block, are suspicious of relying on cases of blindsight and neglect to defend unconscious perception because they involve damage resulting in neuro-atypicality. It's difficult to tell in these cases if the damage has truly resulted in a lack of conscious awareness, as opposed to full-blown conscious awareness coupled with an inability to report.[12] Block prefers to defend cases of unconscious perception from our second category. These involve neurotypical individuals whose perceptual contents are purportedly rendered unconscious

[11] Marshall and Halligan 1988, 766, see also Phillips 2015, 15.

[12] One might also find it suspicious if cases of unconscious perception were possible only in patients with brain damage. Here, we are minimally exploring the possibility of unconscious content, not pursuing the stronger claim that it is normal or even frequent. In any case, other instances of purported cases involve neurotypical subjects.

due to suppression techniques, including masking, binocular rivalry, and continuous flash suppression (Phillips and Block 2017, 174).

In cases of continuous flash suppression, the two eyes are presented with different stimuli, one of which is a flashing, high-contrast image called a "Mondrian" and the other of which is a normal image of, say, a house or face. The flashing image allegedly suppresses the ordinary image from the subject's conscious awareness, resulting in subjects reporting near total unawareness of it. Nevertheless, evidence suggests some processing of the content of the suppressed image must be taking place. Block cites studies by Jiang et al. (2007) where images of nude bodies presented to subjects appeared to be suppressed from conscious awareness by Mondrians. When subjects were asked questions about the location of the nude body image, they were at chance in responding, suggesting the images were unconscious. Still, after viewing these images, the subjects were faster at identifying properties of the next stimulus provided it appeared on the same side as the nude image and the nude matched the gender that was desirable given that subject's reported sexual orientation (a point that will be important later in the debate).[13] Again, this seems to indicate that the subjects perceptually represented the nude images, despite purportedly lacking conscious awareness of them.

In both of these cases, we have evidence for visual perceptual representation in the absence of consciousness. Critics of these cases, most notably Phillips, object to them on one of two grounds, presenting a dilemma for the proponent of unconscious perception. The first objection is that these findings are arguably not the result of unconscious mental states (as opposed to conscious mental states that the subject fails to report due to personal biases in their responses); while the second objection is that these findings arguably lack genuine perception (as opposed to non-perceptual, mere information processing). The first Phillips labels 'the problem of the criterion' and the second, 'the problem of attribution'. (It is from these more narrow characterizations that I derived the more general characterizations of the problems defined above.)

3.3 Is perception really unconscious?

The first critique of purported cases of unconscious perception is that they are not clearly cases lacking consciousness. It seems our best evidence as to whether some state is conscious or unconscious for some individual is through the individual's own verbal reports. As Block

[13] Phillips and Block 2017, 177-78.

(2007, 535) states, “introspective reports do have a certain priority: we have no choice but to start with reports in investigating consciousness.” This makes sense since consciousness is ultimately a matter of subjective experience, and who better to report those experiences than the subject. However, notoriously, such reports are also subject to response bias. Recall the questions asked in the neglect study:

1) Are the houses the same or different?
2) Is there anything wrong with one of them?
3) Which would you want to live in?

The worry is that some of these questions might prompt conservative responses in light of, say, uncertainty or lack of confidence, which themselves could be the result of degraded or fragmented conscious perception rather than from unconscious perception. In these cases, we might say that the subject is only “dimly conscious” of the relevant stimulus, and that conscious awareness is so dim that they hesitate to rely on it when producing subjective reports.[14]

We can imagine that P.S. does have *some* conscious awareness of what’s occurring on the left side of her visual field, it’s just that this awareness is weak, perhaps even incomplete—she is merely “dimly conscious” of it.[15] This would give her an inkling that something is different between the two houses, but she might be hesitant to trust that inkling. Thus, in response to questions one and two, she might decide not to trust her inkling, instead preferring to err conservatively on the side of reporting no difference. One can imagine her reasoning that if, say in response to question one, she reports the houses are different, then she might then be on the hook for saying in what way they differ—a question she doesn’t feel confident she can provide a precise answer to. However, when confronted with question three, it’s not clear which option would be more conservative

[14] Response bias can be due to many factors outside of mere uncertainty. It might be that subjects are aware of what the experiment is intended to study, and so (wittingly or unwittingly) conform their responses to them; or they might simply have other personal motivations for responding in particular ways, like to preserve certain self-conceptions. These other ways response bias thwarts measures of consciousness will be especially relevant in the discussion of unconscious social bias.

[15] Likewise, blindsight subjects often state that although they cannot precisely identify the stimulus in front of them, they are aware that something is happening. Consider, for example, what’s often called ‘type-2 blindsight’, which differs from ‘type-1 blindsight’ in that patients with the former but not the latter “under some circumstances, report some kind of experiences associated with stimuli presented in their regions of blindness” (Kentridge 2015, 41). These cases are typically instances wherein blindsight patients can detect sharp changes in contrast or motion, but are not able to discriminate what sorts of changes they’re detecting. In other words, they have a feeling that something happened, but they can’t say what it is.

than the other. No matter what she chooses, she can always fall back on claiming she just picked one at random. Therefore, her responses to this question seem free from response bias and are taken to more accurately reflect her conscious experience. If this story is right, then what appeared evidence for unconscious perception can be explained away by response bias.

Phillips puts these points in terms of Signal Detection Theory (SDT). Here, I summarize the theory in order to highlight two advantages it provides in engaging with the conscious criterion problem. First, SDT has the advantage of distinguishing between measures of tasks that are and are not subject to response bias. Returning again to P.S., her ability to discriminate between the houses revealed by question three is taken to be free of response bias. In SDT, this ability is measured by d', where a value of d' greater than zero indicates some ability to discriminate between signal and noise. That said, her willingness to report same/different or yes/no revealed by questions one and two was taken to be subject to response bias. Thus, it is a product of some combination of d' and her particular subjective response criterion c. Think of c as a threshold set by each subject (not necessarily deliberatively) for responding. If the subject has some discriminative capacity (d' $>$ 0) *and* that capacity has exceeded their personal threshold (c), then their reports are likely to reflect their discriminative capacity. However, it's possible that they have some discriminative capacity (d' $>$ 0) that falls below their personal threshold c, in which case their reports will not reveal the full range of their capacity to detect signal from noise. Looking at extremes, one can imagine a subject that has an extremely conservative response bias (say they, for whatever reason, refuse to ever report seeing a stimulus), which would result in learning almost nothing about their discriminative capacities.

The second advantage of SDT, related to the first, is that it helps keep track of four possibilities that can occur for any discriminative task involving a combination of signal and noise: first, a subject can say there's a signal when there is (a true positive); second, they can say there's no signal when there isn't (a true negative); third, they can say there's a signal when there isn't (a false positive); and finally, they can say there's no signal when there is (a false negative). Notice how these might intersect in interesting ways with c. If there is never a signal present and the subject has an extremely conservative response bias, they might accidentally register a lot of true negatives. Likewise, if there's always a signal present and the subject has an extremely liberal response bias, they might accidentally register a lot of true positives. Both would give the appearance that the subject is quite accurate in their discriminative abilities even if they're not; were the signal to fluctuate, we would see that they have many false positives and false negatives, respectively. Thus, the base-rates of the stimulus are important for measuring overall accuracy.[16] Reading discriminative capacity off of results becomes even more complicated if we allow for the possibility of subjects changing their response criterion c from trial to trial.

In sum, some observed behaviors that seemed evidence of unconscious perception might not in fact be tracking conscious awareness. This is because those measures track behaviors that are a combination of many features, including subjective response criteria. All behavioral responses that are not free of potential response bias—which includes all subjective reports about what a subject sees—will be subject to what we (partly following Phillips) call "the conscious criterion problem". In all cases we can ask: is this person really not conscious of what they see, or are they merely not confident enough to report it to others?

What's important for our purposes is that progress can be made on this issue by relying on sophisticated (psychophysical) metrics that can tease apart objective and subjective criterion influences. According to Phillips, SDT is one metric that shows promise in this domain. In the end, Phillips and Block agree that there are empirical tests where sophisticated metrics have arguably been used to measure objective discrimination capacities that operate below conscious awareness. However, Phillips will respond that in all such cases where we have a good candidate for unconsciousness using such methods, such cases will not involve genuine attributable content. They lack the relevant conditions that would distinguish these cases from mere low-level, non-representational information processing. In other worse, they fall victim to the content attribution problem, which I turn to next.

3.4 Is unconscious perception really attributable content?

According to Phillips, the attribution problem distills into a problem about individual-level attributability. Attributability to the individual will, for him, require some connection to voluntary control. On this, Phillips at various times contrasts, on the one hand, states and processes that are attributable to the individual by dint of being "available to central agency", being within "voluntary, agentive control", being those that "the subjects can themselves use", being those that can be "exploited by subjects to guide and control their actions", and being those that operate in accordance with "subject's knowledge and intentions"; while on the other hand, states that are not attributable to the individual by dint of being "completely stimulus-driven reflexes", of operating "entirely outside of

[16] This point will be important later when discussing recent evidence for unconscious social bias.

voluntary control", and outside of "direct control."[17] Crucially, according to Phillips, all of the cases of unconscious perception that do not fall victim to the criterion problem (and some that do), fall victim to this attribution problem, because they fail to meet these standards for attribution to the individual.[18]

Block, like Burge, denies that availability to central agency is a necessary condition for attributability to the individual. Likewise, he argues that his preferred cases of unconscious perception are attributable to the individual, since the representational contents they involve are high-level contents connected to whole-individual aims. Consider again the Jiang et al. (2007) study. Because the nude images only primed identification in cases where they matched the sexual orientation of the subject, the study suggests that the subjects were registering high-level content (the gender of the individual in the photo) and that that content was connected in appropriate ways to desires and preferences of the individual. Phillips disagrees, stating among a variety of reasons that it is not the contents of a state that determine its attributability, but rather its role in the guidance of voluntary control (which he argues is a role the states in Jiang et al. lack).[19]

This is just a small sampling of a rich discussion regarding the relationship between action, behavior, agency, and attributability to the individual.[20] There's a lot more to say about the notion both within and independent of the debate about unconscious perception. Both Burge and Phillips acknowledge that these concepts need further development than is provided by either of their works.[21]

I believe it was a mistake to place so much weight on the notion of attributability to an individual within Burge's characterization of perception. Contrary to what Phillips

[17] For a sampling of these phrases, see Phillips and Block 2017, 181 and Phillips 2018, 495, 496, 498, 499.

[18] One obvious worry with this approach is that, if you think voluntary control is synonymous with *conscious* voluntary control, then we're back to a notion of perception that presupposes consciousness. Thus, this conception of individual attributability runs the risk of, as Robert Kendridge warns, "slipping in a requirement equivalent to 'perception must be conscious' through the back door" (Peters et al. 2017, 4).

[19] See Phillips and Block 2017, 173-174 for discussion.

[20] The notion of attributability to the individual is intended to correspond to a more general distinction that is echoed in work far beyond Burge's. There have been, throughout the history of philosophy of mind, many attempts to characterize a distinction between two levels of states within a creature's psychology. For Burge (2010, 2022), the distinction is between being by the individual and not being by the individual. For Dennett (1969), the distinction was between being personal and being subpersonal. For Sellars (1956), it was belonging to the space of reasons vs the space of causes. For Stich (1978), it was being doxastic vs subdoxastic. For Fodor (1983), it was being modular vs belonging to central cognitive capacities. As Block says, all such characterizations come with an air of postulation (Peters et al. 2017, 8).

[21] Burge 2010, 335, fn. 62, Phillips 2018, 499, fn. 45. For impressive work dedicated to the development of these concepts, see Buehler 2014.

reports, attributability to the individual is not a necessary feature of *all* perception according to Burge.[22] While it's true that Burge states, "fundamentally, it is the individual that perceives," he prefaces that statement with "I do not claim that *all* perceptions are perceptions by an individual."[23]

Phillips (and Block) may diverge from Burge on this point too. Their doing so will seemingly bring us back to a dialectical impasse, unless we can find some independent motivation for a neutral characterization of the attributability requirement. I take it that the most compelling reasons to prefer an attributability requirement is that, in the absence of it, there would be no clear criteria for calling something perception. Phillips often makes statements like this when responding to claims that coordinated central agency is not required for perception (Phillips and Block 2017, 181). In the absence of this feature, what positive reason could we provide for regarding some state as perception? Echoing this concern, in discussion of why individual attributability is important, Phillips (2018, 498) quotes psychologists Klotz and Neumann (1999, 976), who reason the following:

> In ordinary usage, perceiving is something that a person or animal does, not something that can be properly ascribed to stages, subsystems, brain areas, or the like. The triggering of a sneeze by an external stimulus does not imply that the reflex center that controls it 'perceives' the stimulus.[24]

It seems then the concern boils down to this: in the absence of clear criteria for individual attributability, there will be no way to distinguish clear cases of non-perception, like the brute causal and transduction mechanisms of the brain, from genuine perception. Put another way, there will be no way to distinguish the kind of robust intentional content we were interested in as a mark of the mental from the more superficial statistical notions of aboutness demonstrated by the rock and shadow.[25]

[22] Phillips (2018, 481, fn. 19) states that he, Block, and Burge all agree that attributability to the individual is a necessary feature of perception. On this, he says that while Burge maintains that some "perceptual representations" might not be attributable to the individual, perception always (and constitutively) is. Here, Phillips is resting on a distinction made by Burge between perceptual representations and perception proper. It is true that Burge will make this distinction (recall in my discussion of the necessity and sufficiency of the four elements of Burge's characterization that perceptual anticipation and memory are perceptual but not perception). But he does not do so when stating his view that perception need not be by the individual.

[23] Burge 2010, 369, emphasis in original. This point is even more explicit in Burge's direct criticisms of Phillips's misinterpretation of his view. See Burge 2020, fn. 45 and Burge 2022, 19.

[24] In similar spirit, Quilty-Dunn (2019, 462) states, "mere sensory processing of this sort [transduction] is arguably insufficient for genuine perception" (citing in that context Prinz 2015 and Phillips 2015).

[25] This motivation is related to what Krickel (2022) describes as "the unconscious mind worry": in the

But distinguishing between brute causal mechanisms of sensory registrations and genuine representational states is precisely what the other criteria of Burge's characterization do. We already have other resources required to overcome this challenge, without resorting to controversial criteria of individual-level attributability. By focusing instead on the attribution of content, understood as some states being both *objective* and *representational* (grounded in scientists' attributions of constancy mechanisms), we have a clear criterion for distinguishing states that are genuine perception from mere sensory registrations. In fact, I think we can bolster this line of reasoning by ruminating on what importance perception's being by the individual was initially intended to capture within Burge's theory. This will simultaneously give us a more nuanced understanding of the content attribution problem and clarify strategies for overcoming it in particular cases.

What's important in perception's being fundamentally (though not always) attributable to the individual, as discussed by Burge (2010, 370), is that the *kinds* that are eligible to be the objects of representational states are tied to whole-individual function, e.g., eating, mating, flourishing. Going back to the berry example, it's important that we're able to represent distal features of the environment, like the berry and its constant reflectance, because this capacity contributes to whole-individual function. It is, for example, berries that are edible or inedible (not light just before it hits the retina), and it is the berries' reflectances that give us clues about their being edible or inedible. These distal features (and our ability to represent them objectively) are what factor into whole-individual functioning. This isn't to say that representational function can be reduced to biological function.[26] However, it is the case that whole-individual function helps to delineate what kinds can factor into representational capacities. Some kind's being a candidate for representational content (as evidenced by our ability to represent those kinds) is enough to link it up in significant

absence of consciousness, what reason do we have for regarding some state mental (as opposed to merely biological). Likewise comparing the two debates of unconscious perception and unconscious bias, she offers a mechanistic response wherein what determines whether some process is mental turns on to what extent some mechanistic difference makers are present in explanations of some behavior in both conscious and unconscious instances. Insofar as both her and my view turn on how some state features in explanations, the two views are congenial to one another. As a biographical note, Krickel?s paper was published as I was finishing the final edits of this paper. I regret not being able to offer a more sustained engagement with it here. No doubt the present paper would have benefited greatly had I encountered it earlier in the writing process.

[26] Burge (2010, 292-308) goes to great lengths to distinguish between mere biological functions and representational functions. Likewise, not all whole-individual function (even apart from representational function) will be biological and in principle creatures without biological function (like robots) might still have perception (Burge 2022, 26; see also his fn. 46 and p. 280).

and important ways to the goals and desires of the individual. Crucially, none of this depends on overly demanding notions of individual actions, such as that they be within deliberate control of the individual or that they are available to consciousness. Rather, these considerations bring us back to central notions of objectification and a state's having the representational content that it does.

As this discussion brings out, our impetus for content attribution will be tied in fundamental ways to the *kinds* that factor into a state's veridicality conditions. Regarding a state as genuinely contentful (as opposed to reducing it to mere causal, non-representational information processing) will depend on the kinds and properties we think are being tracked by the information processing capacity. When an explanation of this capacity's operation is restricted to properties and kinds instantiated by the proximal stimulus, this will be reason to avoid attributability of genuine representational content. This is why we don't attribute content in the case of sensory registrations: we can make sense of their operation by only ever citing properties of the proximal stimulus. However, when in an explanation of some capacity we must make reference to robust distal kinds, those that are tied in fundamental ways to whole-individual function, then that will be reason to attribute genuine representational content. This also explains why constancy mechanisms are such good hallmarks of genuine perception: these just are capacities that allow us to track constant distal kinds in the face of ever-changing proximal stimulations. This is not something we can fully explain without making reference to the distal kinds themselves.

I believe this is a better condition of content attribution than one tied to individual-attributability as offered by Phillips. What's special about perception is that its content is what it is in virtue of objective, representational capacities that are fundamentally tied to individual-level function. Those contents are what they are (pick out the kinds that they do) in virtue of this connection. Constancy mechanisms are the hallmark of these representations (and misrepresentations) of distal features of the environment. They enable us to get beyond the veil of appearances to objective representation of the world. This is philosophically central.

If this is one's interpretation of the attributability problem, then many cases of purportedly unconscious perception will avoid the content attribution problem. For example, the Jiang et al. (2007) cases cited by Block appear to meet this condition for the reasons stated by Block: that the content is tied in fundamental ways to individual-level pursuits. Likewise, for Burge, many instances of blindsight involve grasping behavior explained only by reference to shape properties of the objects being grasped (i.e., distal properties). Even

Phillips admits of the existence of empirical priming studies that satisfy the more rigorous demands of the conscious criterion problem (demonstrated by the employment of SDT) and that involve constancy mechanisms, thereby satisfying the demands of this refined content attributability problem.[27]

3.5 Summary of Insights

This review of the debate about unconscious perception has highlighted two problems any purported case of unconscious content must address, while simultaneously outlining strategies for addressing each. The problems are the conscious criterion problem and the content attribution problem. The first demands that we distinguish purported cases of unconscious content as those that are truly unconscious, as measured by some objective conscious criterion, and not merely the byproduct of interaction effects with some subjective response criterion. To address this issue, empirical work can adopt sophisticated (psychophysical) metrics such as SDT that can adjudicate effects of the two. The second demands that we distinguish purported cases of unconscious content as those that truly attribute content, as measured by capacities that track distal features of the environment, and not merely causal, non-representational information processing. To address this issue, empirical work can carefully delineate those cases where their attributions of content essentially involve robust distal kinds and properties that feature fundamentally in whole-individual function (such as when visual psychologists posit constancy mechanisms). These cases cannot be reduced to mere causal descriptions involving low-level properties of proximal stimulations.

These will serve as extendable insights into the debate about social bias, which I turn to next.

4 Is unconscious bias unconscious bias?

Debate about unconscious bias, though still in its infancy, mirrors many aspects of the preceding debate about unconscious perception.[28] Following the path laid out by that

[27] Phillips offers Norman et al. 2014 as such a case, though he ultimately denies it as a case of unconscious perception since, according to him, it does not meet the individual-attributability requirement. See Phillips and Block 2017, 168-169 for discussion.

[28] Often scholarship on unconscious bias calls them "implicit bias" or "implicit attitudes". The word 'implicit' is used in many ways in the literature. For many, it means unconscious. For others, it is associated with other properties themselves associated with so-called 'system 1' processes, e.g., being fast, automatic, and subpersonal. More recently, there's been a push in the literature on implicit bias to use 'implicit' to

debate, I'll address first whether we can characterize bias in a consciousness-neutral way; second, whether there are instances of bias that are unconscious (keeping in mind the consciousness criterion problem); and third, whether there are instances of bias that involve genuine attributable content (keeping in mind the content attribution problem).

4.1 What is bias?

As in the case of unconscious perception, for the question of whether unconscious bias is possible to make sense, we need some notion of bias that does not presuppose conscious accessibility. In my work, I've argued for a general, functional account of bias.[29] In a functional account of bias, we specify what bias is by the functional role that it serves: roughly, by its input-output behavior or, rougher still, the ways that it guides transitions between informational states.[30]

According to this account, bias exhibits a functional profile (that is, it interacts with other mental states in ways) that mimics inductions on the basis of environmental regularities. Part of what defines bias's functional role is its response to underdetermination present in induction. Bias originates for the purpose of overcoming this underdetermination, and it does so by tracking environmental regularities. Thus, bias is embodied in the reality-tracking rules that guide induction. In short, biases are assumptions (heuristics, tendencies, norms) that allow us to limit the inductive hypothesis space to a tractable size by assuming what is normal in our environment.[31]

Consider again the discussion of perceiving the red berry above. This case illustrates a perceptual bias. Because the sensory registration of light underdetermines the possible distal objects that can cause the sensory registration, it also underdetermines the possible representational contents that can be formed on the basis of that registration. In order to overcome this underdetermination, the visual system adopts the bias to subtract surround-

describe the indirect tests for mental states (like the IAT) rather than the mental states themselves. In my own work, I use it more aligned with philosophical theories of representational content, where 'implicit' is taken to mean not explicitly represented, i.e., merely encoded in the operation of the computational system. Thus, I claim that some "implicit biases" are "truly implicit". See Johnson 2020c, 1201-05 for discussion. Here, I will as much as possible use "unconscious bias" rather than "implicit bias" to avoid further confusion.

[29] See Johnson 2020a, Johnson 2020b, and Johnson 2020c.

[30] This is opposed to an account that would specify what bias is by, for instance, a specific neural mechanism in the brain, a particular representational content or format, or by a simple behavioral disposition.

[31] In its focus on induction and underdetermination, my account owes much intellectual debt to Louise Antony's work on bias (Antony 2001, 2016).

ing light. This bias, together with the sensory input, allows the visual system to output a representation of a red berry. In this way, the assumptions embedded in constancy mechanisms throughout the visual system are all biases that encode regularities of one's normal physical environment (e.g., that illumination is uniform, that light comes from above, that patterns are homogenous). When a person has a perceptual bias, their visual system has built-in constancy mechanisms that take them from underdetermining 2-D informational states to 3-D representations of the world.

Crucially, as previously discussed, the biases illustrated in cases of visual perception are not realized by states with representational content that is attributable to the individual. The assumptions encoded in these transformation processes are not obviously explicitly represented, but arguably tacitly built into how the computational processes operate.[32] However, the flexibility of the functional account allows for variation here. Crucially, the functional account remains agnostic about what sorts of mental entities bridge the gap between the underdetermining inputs and outputs, ultimately highlighting the diversity of candidates that can serve the role and allowing those that are genuinely contentful. These differences can be seen by extending the notion of bias to cases of social bias.[33]

Social bias, like perceptual biases, facilitate inductions made on the basis of regularities within the environment. However, in the case of social induction, the relevant features of the environment will be social groups and the properties taken to be prevalent among their members. When a person has a social bias, they have some combination of states and processes that take them from underdetermining inputs (in the form of beliefs that someone belongs to a social group) to outputs (in the form of beliefs that that person has some characteristics stereotypical of the social group). Social biases, then, encode assumptions about which properties are stereotypical of a social group (e.g., a stereotype that elderly people are bad with computers).[34] Crucially, these states and processes that

[32] See Johnson 2020c for a full defense of this point.

[33] This general notion of bias can likewise be extended to the biases that exist in machine learning programs, where just as with psychological biases (perceptual or social), algorithmic biases mimic inductions made on the basis of regularities in the environment too. These algorithms imbibe regularities in the environment (e.g., that people type 'male' after typing 'doctor', that images of a particular luminance correspond to depth, that people raised in single-parent households are more likely to recidivate) and use them to inductively label new data points. See Johnson 2020a and Johnson 2020b for more discussion.

[34] Here, I'm using 'stereotypical' in a neutral way to include social stereotypes as we normally understand them, but to also include broad regularities that are not taken to be social. To draw an analogy with the visual perceptual case, we can speak loosely about the "stereotypical" conditions wherein illumination is uniform. Thus, perceptual biases, like social biases, function to track stereotypes loosely speaking. Of course, they could ultimately fail in this function, i.e., some regularity assumed by the system might not

realize the bias can take many forms. They could be a generic belief, some complex association between solitary concepts, or some other mental construct altogether. In order to walk through how different elements might correspond to our notion of *bias*, it will help to again have a concrete (though idealized) example.[35]

Imagine a scenario where a person 'E' has a bias against the elderly that causes him to negatively evaluate elderly individuals he encounters. In this case, let's imagine E runs into his fellow colleague Jan, and forms a negative evaluation of her, i.e., he dislikes Jan. When asked why he dislikes Jan, he explains that it's because Jan is elderly and that he dislikes elderly people in general. Here, E has an explicit bias, and the inference he's making seems straightforward:

(*i*) Jan is elderly.
(*ii*) I dislike elderly people.
∴ (*iii*) I dislike Jan.

Although it's clear there's a bias here, it's less clear which mental states and actions correspond to *the bias*.[36] Just focusing on the mental states involved, the bias could be the generalizing belief corresponding to (*ii*) or the specific conclusion he draws about Jan, which corresponds to (*iii*). The functional account allows us to distinguish between these. On the functional account, we call the belief about a particular person on the basis of which a discriminatory judgment is formed—in this case, E's belief that Jan is elderly—*the bias-input*. Next, we call the collection of states and processes that—in tandem with the bias-input—cause a discriminatory judgment *the bias-construct*; the bias-construct in

obtain. A system might assume some loose regularity between the properties *being elderly* and *being bad with computers*, but it might turn out that no such regularity actually exists. Moreover, I'm not intending to beg any questions against different views about what is distinctive about social stereotypes. For sophisticated discussion of this question, see Beeghly 2015.

[35] Much of the example and discussion to follow is borrowed (with slight modifications) from Johnson 2020c, 1196.

[36] The word 'bias' is often ambiguous. We could refer to someone's acting in discriminatory ways toward members of social groups as their 'bias'. This would suggest a more behavioral interpretation of bias, one that is becoming increasingly popular in the literature (see, for example, De Houwer 2019). Gawronski et al. (2006) famously distinguished between three different focal points in discussions of social bias awareness: source awareness, content awareness, and impact awareness. These distinctions have now become standard when discussing in what ways, if any, social bias might be unconscious. In fact, these distinctions are not between different types of awareness, but rather different objects that could be the target of awareness. They correspond roughly to the environmental and psychological causes of someone's acquiring a bias (source), the bias itself (content), and the behaviors and psychological states that result from the bias (impact). For reasons to follow, there are important distinctions to be made among the mental states and processes that are seemingly collapsed under Gawronski et al. (2006)'s notion of 'content'.

E's case is his generalizing belief that he dislikes elderly people (together with whatever inferential processes are necessary to derive the conclusion). Additionally, we call the discriminatory judgment that bias-constructs and bias-inputs together cause—like E's belief that he dislikes Jan—*the bias-output*. Finally, we can call actions that are performed on the basis of bias-outputs—like E's avoiding Jan in the hallway—*bias-actions*.[37] 'Bias' simpliciter, then, is reserved for bias-constructs. They're what lead us from bias-inputs to bias-outputs, resulting in bias-actions.

In E's case, his bias (bias-construct) is instantiated by an explicit belief. Thus, the bias takes the form of a consciously accessible and fully attributable representational content. However, it needn't be the case that all biases take this form. It might be that in other cases, the bias-construct is either unconscious or not constituted by representational content attributable to him. The functional account leaves room for flexible bias-constructs, wherein many different states and processes can map the same kinds of inputs to the same kinds of outputs.[38] In this way, all biases can be characterized along the following tripartite, functional model:

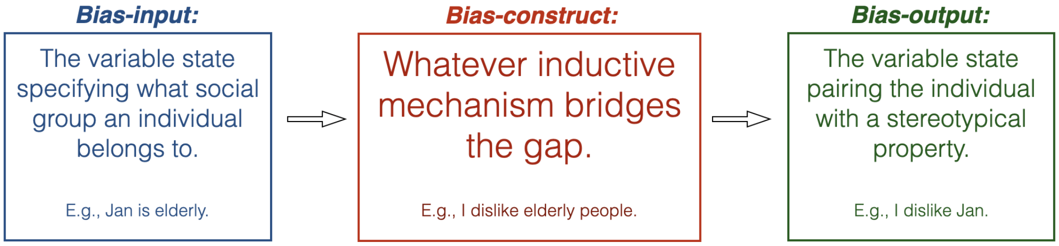

Bias-input:
The variable state specifying what social group an individual belongs to.
E.g., Jan is elderly.
Bias-construct:
Whatever inductive mechanism bridges the gap.
E.g., I dislike elderly people.
Bias-output:
The variable state pairing the individual with a stereotypical property.
E.g., I dislike Jan.

Figure Three: The Functional Model of Social Bias.

The debate about unconscious social bias to follow, then, will center around what sorts of states and processes constitute the bias-construct, whether those states and processes are truly unconscious, or whether they're bonafide attributable representational contents.

[37] A similar three-part functional distinction can be made in the case of stereotype implicit biases, as I originally discuss in Johnson 2020c. However, since the data on the conscious accessibility of bias deal almost exclusively with evaluative implicit biases, I discuss only that case here.

[38] A point I argue for at length in Johnson 2020c.

4.2 What is unconscious bias?

In its early stages, literature exploring the existence of so-called "implicit" biases often characterized them as unconscious.[39] In their landmark paper, Greenwald and Banaji (1995, 8) define 'implicit attitudes' as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects." On this characterization, "implicit biases" are arguably attitudes that are unconscious.[40]

In the standard cases, evidence of unconscious social bias comes from comparing results on two types of tests: direct and indirect. Direct tests involve asking subjects to report their social attitudes toward others. Social attitudes can be split into either evaluative attitudes (roughly, warm or cold feelings toward members of social groups) and stereotypical attitudes (roughly, whether they think some properties are stereotypically associated with members of social groups). Some of the most standard measures for evaluative attitudes are the Modern Racism Scale (where subjects are asked questions thought to reveal evaluative social attitudes, like 'Discrimination against [social group X] is no longer a significant problem in the United States') and a "feelings thermometer" (where subjects are asked to rate their feelings toward members of particular social groups on a range of 0 "cold and unfavorable feelings" to 100 "very warm and favorable feelings"). In the case of direct measures of stereotype attitudes, subjects might be asked directly to what extent they agree with stereotypical statement (e.g., 'do you agree that elderly people are bad with computers?').

Indirect measures are intended to test attitudes without asking subjects to directly report and instead measuring their results on some behavioral task. The most famous test of this sort is the Implicit Association Test (or IAT). In the IAT, subjects are asked to quickly sort stimuli presented to them on a computer screen to the left or to the right. The stimuli fall into four categories, and whether two among the four of those categories are "associated" with one another compared to the other categories is determined by how

[39] This describes at least one prominent strand of the literature. Another strand tended historically to characterize implicit bias in terms of automaticity rather than consciousness. See Gawronski and Payne (2010, 2)'s discussion of "two roots of implicit social cognition." Thanks to Alex Madva for urging me to highlight this alternative historical framing.

[40] In a footnote, Greenwald and Banaji explain the parenthetical of "inaccurately" identified traces as intending to include instances where an individual remembers particular experiences, but is unable to identify how those experience shape further processing, e.g., "a student may be aware of having been graded highly in a course, but not suspect that this experience influences responses to the course's end-of-term course evaluation survey" (p. 8, n. 2).

well a subject can quickly sort them to the same side of the screen. For example, if I wanted to test to what extent a person associated "elderly people" with the stereotypical property of being "bad with computers", I might give them an IAT that has four kinds of stimuli: pictures of elderly people, pictures of non-elderly people, pictures of computer-related objects (like a keyboard), and pictures of non-computer-related objects (like a legal pad). If when asked to sort the pictures of elderly people and non-computer-related objects to the same side of the screen (sorting the other objects to the other side), they are faster and make fewer mistakes than when I ask them to sort elderly people and computer-related objects to the same side of the screen (sorting the other objects to the other side); then this is taken as good evidence that they harbor some bias that pairs the social group *elderly* with the property of *bad with computers*.

Evidence for unconscious bias then comes in two varieties: first, divergent results on the two kinds of tests (direct and indirect); and second, surprise and shock when being confronted with results on indirect tests. Diverging results on the two kinds of test occurs when, for example, an IAT provides evidence that a subject harbors a social attitude, say a stereotype that elderly people are bad with computers, but yet reports on a direct test that they do not. Imagine that a subject responds 'no' when asked if elderly people are bad with computers, but they're slower at sorting elderly faces with computer-related objects. This is taken as evidence that they have a mental state with the content that elderly people are bad with computers (evidenced by the IAT), but they don't have conscious awareness of that mental state (evidenced by the direct subjective report). Meta-analyses suggest that correlations between results on direct and indirect measures are low in just the sort of way that would suggest divergence (Hofmann et al. 2005, Cameron et al. 2012). Likewise, people often demonstrate surprise and shock when confronted with their indirect test results. This sort of experience can be self-administered by taking an IAT online, but they have also been empirically codified in studies that demonstrate the same point[41]

Taken together, these two sorts of results suggest that people harbor social biases that they are not aware of. But, as with the case of unconscious perception, this evidence is subject to both the conscious criterion problem and the content attribution problem. In what follows, I discuss each in turn.

[41] Hillard et al. 2013, Howell et al. 2015, and Howell and Ratliff 2017. More precisely, these studies show that strength of pro-white preferences is predictive of negative affect including surprise (Hillard et al. 2013, 506), where most subjects demonstrate pro-white preferences and that, in general, "people responded defensively to feedback indicating they were biased" (Howell and Ratliff 2017, 125).

4.3 Is unconscious bias really unconscious?

Evidence against the claim that the above reported social attitudes are really unconscious come either as critiques of the evidence cited above or, more recently, positive evidence that subjects can in fact report the attitudes once thought unreportable.

Starting with the critiques of the above methodology, perhaps the most looming concern (as with the discussion of unconscious perception above) is response bias. In the aforementioned cases, it might be that although subjects report not harboring social biases, they are perfectly aware that they do, they just don't want to admit it.[42] But we needn't assume that subjects are purposely trying to deceive experimenters. It might simply be that (again, as with unconscious perception) they are only weakly aware of their social biases, and their subjective response criterion is conservative enough that they do not feel confident in reporting these attitudes when asked. Indeed, divergences are minimized in conditions where subjects additionally report low motivations to control prejudice (e.g., Dunton and Fazio 1997), suggesting quite liberal response criteria can mitigate divergence.

There are also concerns that the divergence between direct and indirect measures are due to a variety of other methodological inconsistencies.[43] For example, often experimenters, wanting to avoid exactly the kind of response bias discussed above, use questions in direct measures that are not so obviously about the subjects' evaluative attitudes. For example, questions from direct measures like the Modern Racism Scale tend to focus on evaluative attitudes concerning, in addition to the social groups in question, some governmental policy or other (e.g., affirmative action). Thus, reports in response to these questions will be a combination of the subject's evaluation of both the social group and their perception of whether the policy is valid. Another concern is that often direct and indirect measures are asking questions about members of social groups at different levels of granularity. Whereas direct measures typically ask subjects to report on group-level, generic attitudes ('do you think elderly people are bad with computers?') the behavioral tasks measured by indirect tests are almost always at the individual-member level (sorting

[42] There's reason to think response bias (and, thus, the criterion problem) will be significantly harder to overcome in the domain of social psychology. Subjects likely have a more vested interest in hiding their socially unacceptable prejudices than whether they see some innocuous visual stimulus. Indeed, it was precisely the concern that individuals harbored negative social biases that they didn't want to admit to that prompted the shift in empirical psychology from direct to indirect measures in the first place (Banaji and Greenwald 2013, 170-184). It's hard to see how, then, empirical methods in this domain could ever fully avoid conscious criterion concerns.

[43] See Gawronski 2019, 578-580 and Hahn and Gawronski 2014, 28-29 for careful discussion of the points that follow.

the faces of particular elderly people). And finally, measures of social attitudes have been demonstrated to exhibit wide contextual effects, such that direct and indirect measures differing in context (either the context embedded in the tests or the subject's context when taking the tests) might limit the sorts of conclusions one can draw about their results. The thought is that in all of these cases, until we better calibrate direct and indirect tests to be targeting the very same attitudes (with the very same types of representational contents), there's no reason to think that divergence between the tests is indicative of unconscious social bias.

The evidence for shock and surprise is similarly explained away by methodological mismatches. Results on indirect tests like the IAT are typically conveyed using a scale that compares the individual taking the test to the wider population.[44] For example, the IAT results are reported as 'slight preference', 'moderate preference', or 'strong preference' for one group over another as compared to the general population. When a subject is confronted with the result that their IAT demonstrates a "strong" bias, they might be surprised and shocked not because they were unaware of the bias itself, but because they wouldn't have regarded it as strong or because they took themselves to be less biased than those around them.[45]

In addition to the arguments intended to undermine evidence against the conscious accessibility of bias, there's also been a recent surge in literature that purports to give positive evidence in favor of consciously accessible. By and large, these data come in the form of experiments designed to test how well subjects can predict their results on indirect measures like the IAT. For example, Monteith et al. (2001) demonstrate that individuals can accurately describe how they did on an IAT before their results are shared with them.[46] Crucially, it's possible that in these cases, individuals are not introspecting their attitudes, but rather inferring to the existence of such attitudes on the basis of behavioral data they gather from themselves while taking the test. In this case, their ability to predict would be no more evidence of their own conscious awareness than the experimenter's ability to

[44] See Hahn and Gawronski 2014, 29, Hahn and Gawronski 2019, and Hahn and Goedderz 2020 for careful discussion of the points that follow.

[45] While compelling, this explanation would seem to struggle to account for those cases where subjects are surprised by even the slightest result indicating bias on the test. Consider, for example, the surprise demonstrated by individuals who receive a 'slight preference' result on an IAT when they belong to the very same social group that they demonstrate a bias against (see Banaji and Greenwald 2000 for an example of such a reaction). Still, an even more obvious explanation might simply be that people feel bad when told they harbor problematic biases, and so cognitive dissonance compels them to deny it.

[46] See also Hahn et al. (2014) and Rivers and Hahn (2018).

infer on the basis of their behavioral results. In order to address this weakness, Hahn and Gawronski (2019) ran a series of studies testing whether subjects could predict their results before taking an IAT, sometimes merely instructing subjects to direct their attention to their "gut feelings" or "spontaneous reactions" to IAT stimuli (which they provided to them). In these cases too, subjects seemed able to predict whether the test would indicate bias against members of particular social groups. Likewise, after being asked to predict their results while attending to their gut feelings, they reported on a direct feelings thermometer (administered before taking the IAT) results more aligned with their IAT results. These results all seem to indicate that, in fact, subjects do have conscious awareness of the biases measured by the IAT. As Hahn and Gawronski (2019, 24) state, "the current findings are consistent with theories suggesting that implicit evaluations are subjectively experienced as spontaneous affective reactions . . . . Based on these conclusions, we deem it problematic to present implicit biases as attitudes that people are ... unable ... to report."[47]

However, just as response bias loomed large in the critique of evidence against conscious accessibility, so too it looms large here. While it's true that subjective reports became more aligned with their IAT results, it's also true that, on the whole, subjects reported more bias on direct measures compared to before. Taking lessons from SDT discussed above, one might wonder if in fact these results of alignment are demonstrating true sensitivity to subject's internal states, or whether they're the result of subjects merely adopting a different response criterion for prediction. In the face of uncertainty, it's possible that subjects simply preferred to err on the side of saying they had a bias when they didn't, instead of saying they didn't have a bias when they did. In both the Hahn and Gawronski 2019 and similar studies presented by Nier 2005, subjects were arguably primed to expect that they might have attitudes toward social groups that they wouldn't normally express. As the instructions presented to subjects before prediction explained, "you may have a more positive affective reaction toward a picture of a skinny top model than toward a picture of a regular woman, even though you may not think or say that skinny top models are better people than regular women." They were again reminded just before the prediction task that "your first reaction could be different from a general opinion you may have" (Hahn and Gawronski 2019, 789). In the case of the Nier 2005 studies, subjects were shown a clip of an *NBC Dateline* special on the IAT, where throughout, explanations of the test

[47] Such tests also have potential confounds regarding interaction effects between consciousness and memory. I thank Uriah Kriegel for raising this point.

were given that presupposed it would reveal to subjects attitudes that would be surprising to them. It's possible that in light of cues suggesting they might get their predictions wrong, they believe false positives (saying they were biased when they weren't) would be less socially stigmatizing than false negatives (saying they weren't biased when they were), and so erred on the side of the former.[48] As we saw above in the discussion of SDT, this combination of erring on the side of false positives when combined with a scenario where you are more likely to have positive instances (i.e., in cases where biases against social groups are more prevalent, as they tend to be in results on IATs), will merely give the illusion of accuracy. Thus, it's not obvious that direct measures and indirect measures aligning on average is sufficient to indicate conscious awareness. To know for sure would require more detailed analyses of the correlations between direct and indirect studies than are presently available.[49,50]

Thus, evidence for conscious accessibility seems just as victim to the conscious criterion

[48] Indeed, Howell et al. (2013, 716) report that participants of their study "indicated that learning that they were more implicitly biased than they expected would be distressing ... but that learning they were less implicitly biased than they expected would make them happy."

[49] I know of only one study that addresses this potential confound. Nier (2005, 48) states the following:

> [A]nother alternative interpretation that could potentially explain the increased implicit–explicit relationship is that people simply reported more negative attitudes in the Accurate condition, regardless of their level of implicit prejudice. However, the data are not consistent with this interpretation. If all participants in the Accurate condition simply elevated their MRS scores to a similar degree, the relationship between post-test MRS scores and IAT scores would not have been any stronger in the Accurate condition. An increase in the implicit–explicit relationship requires, by definition, that those who had more negative IAT scores reported negative explicit attitudes to a greater degree than those who had more positive IAT scores. Therefore the stronger implicit–explicit relationship observed in the Accurate condition is not consistent with the notion that everyone simply reported more negative MRS scores in the Accurate condition.

The idea here is that in order for correlation between direct and indirect results to increase, not only would those who demonstrated more bias on the IAT need to report more bias on the modern racism scale, but also those who demonstrated no bias (or bias in the opposite direction) on the IAT would need to report no bias on the modern racism scale. What this fails to take into account is that base-rates of bias tend to be, on average, higher than base-rates of no bias. So, in fact, it is possible that everyone on average reporting more bias would indeed raise average correlations. As far as I know, no study provides the more nuanced breakdown of correlations that would be needed to demonstrate that *only* those who demonstrated biases on the IAT predicted higher scores or reported more bias through direct measures.

[50] I do not intend this article as a comprehensive literature review addressing all the data that might speak to conscious accessibility. Rather, I intend it as a small survey of some of the most prominent work cited in the debate. Doubtless other studies exist that investigate manipulations of implicit-explicit correlations and that avoid some of the priming concerns I raise here. Unfortunately, I lack the space to address them in full. I am confident that existing studies are unlikely to fully address both the conscious criterion problem raised here and the content attributability problem raised in the next section. Thanks to Alex Madva for pressing me on this point.

problem as evidence for conscious inaccessibility. But the philosophical debate canvassed above regarding unconscious perception highlights empirical paths to progress here. While it's ultimately within the purview of social psychology to choose precise experimental designs, philosophical reflection suggests that minimally, such designs must adjudicate between objective and subjective criteria effects. Incorporating the lessons from SDT, we need an experimental paradigm that can track objective discriminatory capacities (in subject's reports of their own attitudes) while controlling for subjective response bias combined with high base-rates. One way to do this might be to test for reportability in domains that are likely to have low base-rates of bias.[51] In any case, until empirical methodology is able to disentangle these two effects, it remains subject to the conscious criterion problem. Thus, while I'm happy to allow that for some, the preponderance of evidence will compel them to one side of the debate over the other, I caution against forming any strong conclusions that suggest the empirical evidence is conclusive on the question of whether social bias is unconscious.

Setting this issue of conscious accessibility aside, however, there are likewise independent concerns about whether the evidence in favor of unconscious content really is tracking bonafide attributable representational contents. Thus, we turn to the content attribution problem next.

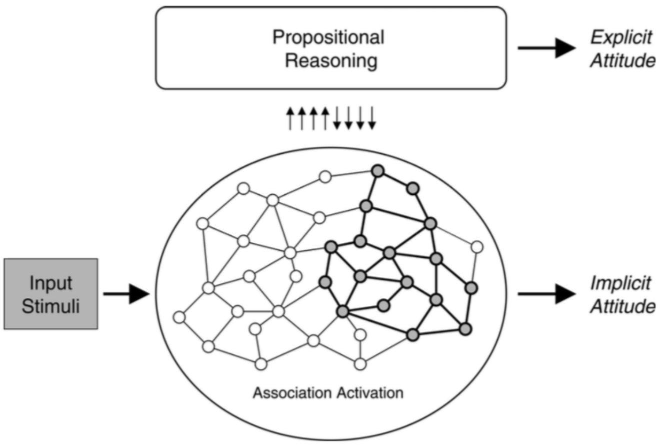
4.4 Is unconscious bias really attributable content?

Let us turn to evaluating the aforementioned empirical studies for the attribution problem. Ultimately, I argue that, even if such studies did indicate that some aspect of a bias's operation is consciously (in)accessible, without being clear about which aspect and what the status of that aspect is as a contentful state, they leave unaddressed the question of whether these are genuine instances of unconscious content. To fully evaluate, we'll need to get even clearer about which aspects constitute the bias.

It helps to work out how the distinctions between components of a bias's operation in the psychological models map onto distinctions made within the functional account described above. The empirical literature above assumes the Associative-Propositional Evaluation (APE) Model of social bias presented by Gawronski and Bodenhausen (2006, 2014). The model suggests that implicit biases and explicit biases involve two distinct

[51] I owe this suggestion to Alex Madva. Though, as he suggests, these domains seem intuitively likely to exhibit high implicit-explicit correlations more generally, generating ceiling effects.

processes: associative processes and propositional processes, respectively. Gawronski and Bodenhausen (2014, 449) hold that the associative processes of implicit biases operate on a network of concepts. These associations then culminate in an evaluative response based on the aggregate valence of the concepts involved in the activation chain. Consider the following figure borrowed from Gawronski and Bodenhausen 2006, 697:



Associtiatve-**P**ropositional **E**valuation (APE) Model

Figure Four: The Associative-Propositional Evaluation Model of Social Bias (borrowed from Gawronski and Bodenhausen 2006, 697).

In this case, both implicit and explicit social biases mimic social kind inductions. This model would explain E's dislike of Jan as an output of propositional reasoning (at the upper-level of the figure). According to the APE model, "propositional processes are defined as the *validation* of the information implied by activated associations, which [is] guided by the principles of cognitive consistency."[52] Essentially, the associative processes go through a process of "propositionalization," resulting in representations that take the form of propositional statements.[53] To return to the example above, E's negative affective reaction toward Jan is the result of the associative network "transformed into propositional statements such as 'X is bad' or 'I dislike X'."[54] This propositional representation is then checked for consistency against the other propositions being considered. If all the proposi-

[52] Gawronski and Bodenhausen 2014, 449, emphasis in original.

[53] Bodenhausen and Gawronski 2013, 958.

[54] Gawronski and Bodenhausen 2014, 450. In this case, it's unclear whether X represents Jan or elderly individuals more generally. For reasons discussed, I charitably interpret them as claiming the latter.

tions are consistent, then the newly-formed proposition is validated. If the propositions are inconsistent, then the inconsistency is resolved by the rejection of one of the propositions.[55]

Implicit biases, on the other hand, do not involve explicit propositional reasoning. For example, say another individual P has a purportedly unconscious bias. P visually perceives Jan who is elderly. This visual perception necessarily activates the mental concept *elderly*. This activation then spreads by way of associative processes to other mental concepts, e.g., *wise*, *frail*, and *forgetful* might all activate. The aggregate valence of these concepts then "elicit[s] a spontaneous affective response that is in line with the net valence of these concepts."[56] In this case, the aggregate valence of the activated concepts is negative, so P would have a negative affective reaction, or what the authors call a "spontaneous affective response". This negative spontaneous affective response is what is then measured by indirect tests. In such a case, the visual perception is the bias-input, the whole associative network then serves as the bias-construct, and the spontaneous affective response serves as the bias-output.

We can now ask whether the Hahn et al. experiments above suggest evidence in favor or against the existence of bonafide attributable content, thereby addressing the content attribution problem. To reiterate the main strategy of the experiments, they attempt to determine whether subjects have conscious access to their biases by asking them to predict their results on indirect methods such as the IAT. Let's say, setting aside criterion concerns, that in general results demonstrate that subjects are fairly accurate in these predictions (even before taking the IAT, so their accuracy cannot be based on bias-actions). Recall that in the Hahn and Gawronski (2019) studies, subjects were asked to predict their results by merely instructing subjects to direct their attention to their "gut feelings" or "spontaneous affective reactions". When doing so, "reactions toward minority members increased acknowledgement of bias to the same extent as IAT score prediction (Study 6), providing further evidence for the functional equivalence of IAT score prediction and attention to spontaneous affective reactions" (p. 23). Here, the researchers are borrowing the notion of "spontaneous affective reaction" from the APE model discussed above, which,

[55] The processes underlying *propositionalization* are never formally laid out by the theory; it says only that the propositions are formed on the basis of the information "implied by activated associations." So, it's unclear how propositions are formed on the basis of the activation networks alone. It's even more unclear when we try to extend the model beyond evaluative biases to stereotype biases, since in those cases there is no valence to aggregate. Thanks to Alex Madva for helpful discussion of this point. See also Madva 2017, 103, fn. 31.

[56] Gawronski and Bodenhausen 2014, 449.

for reasons stated above, should be interpreted as what I've been calling a bias-output, i.e., the result of a bias-construct's operation. Crucially, this means the aforementioned evidence for the conscious accessibility of bias is evidence only that an individual has conscious access to the result of a bias-construct's operation—not that they have access to whatever states and processes mediate the negative affective response.

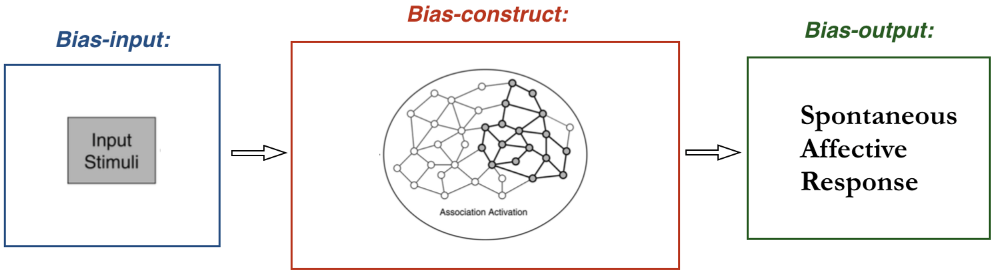


Figure Five: The Combined Associative-Propositional Evaluation Model and Functional Model of Social Bias.

Which should we be concerned about when we evaluate the claim that social bias is unconscious—the bias-output or the bias-construct? Hahn and colleagues seem to be interested in the question of whether a bias-output is consciously accessible. This is in part because they recognize that extent psychological models fail to conceptualize bias in ways that fully distinguish bias-outputs from bias-constructs. On this, Hahn and Goedderz (2020, S127-28) say the following:

> [T]here is a lack of consensus for how the spontaneously activated cognitions ... are best conceptualized. ... To us, it sounds equally plausible to propose that the gut reaction *is* the underlying cognition (and one may then discuss how to conceptualize a gut reaction scientifically), as it would be to propose that the underlying cognition is something else that *produces* a gut response. We hope that continuous advancement in theorizing will shed further light on this question.

This is the perfect opportunity for philosophers of mind to help provide conceptual clarity. On the functional view, bias guides induction. Thinking of social bias in this way helps to secure its place within a broader, unified kind (including other cases of computational bias, like perceptual bias and algorithmic bias). In all the cases of social biases we've discussed (including implicit and explicit biases), the existence of a social bias explains why in the face of underdetermining information about the social group one belonged to, individuals cognitively transition to outputs that paired those individuals with properties stereotypical of the social group. This helps us to explain, for example, why when presented with various individuals belonging to the same social group, the individual is apt to make the same sort of induction. So, while there could be legitimate theoretical interest in the spontaneous affective response (i.e., bias-output), there should also, and perhaps more fundamentally, be theoretical interest in whatever states and processes systematically lead to that response.[57] For any particular response, it seems appropriate to ask why the subject reacts the way they do to particular members of a social group (i.e., what prompts their systematically having the spontaneous affective reactions that they do). And, in ordinary language, it seems perfectly appropriate to respond "because they have a bias against that social group." We then ask whether *that bias* is something the individual was aware of. These sorts of questions seem to me to be at the center of discussions about the conscious accessibility of bias and are necessary to answer for an account of bias to be genuinely explanatory. This places theoretical priority on the bias-construct.

The empirical evidence so far reviewed does little to suggest positive answers to the question of whether the bias-construct is either conscious or contentful. Indeed, they might even serve as unacknowledged evidence for a negative answer to the former.[58] If it turned out that in the standard prediction cases, subjects could *only* predict their IAT results when confronted with concrete stimuli that produced bias-outputs, then that would be a reason to think that they lack conscious accessibility to their bias-constructs. In such cases, subjects must produce bias-outputs in order to have any evidence of the existence of bias-constructs at all. In the absence of such bias-outputs, they lack any evidence whatsoever for the mechanism that systematically produces bias-outputs or bias-actions,

[57] Moreover, as a minor exegetical point, Banaji and Greenwald's original definition of an implicit attitude suggests to me interest in whatever state *gives rise to* the affective response, not the affective response itself: "introspectively unidentified ... traces of past experience *that mediate* favorable or unfavorable feeling, thought, or action toward social objects" (Greenwald and Banaji 1995, 8, emphasis added).

[58] It is important to note that even if subjects are able to predict accurately their results on the IAT, this still is not conclusive evidence that they are accessing representational states that are responsible for their results. As I've discussed elsewhere (Johnson 2020c, fn. 24), a sufficiently reflective thinker might be able to deduce the rules that guide their cognitive transitions. For example, a sufficiently reflective logic student might, upon learning about the rule *Modus Ponens*, reflect on their own personal habits and conclude that they often follow the rule in reasoning. This won't be evidence that that content *as it functioned in producing the inferences themselves* was thereby consciously accessed. Here we might draw a distinction between some content being consciously access**ed** versus being consciously access**ible**. These bias contents might be accessible in some loose sense, but that isn't conclusive evidence that they were directly accessed. Thanks to Uriah Kriegel for raising the importance of this distinction here.

i.e., the bias-construct.[59] As it turns out, this is precisely what emerging data suggest. On this, Hahn and Goedderz (2020, S122) state, "interpretation of the findings presented here suggests that the biases reflected on implicit measures may often be state-unconscious until people are confronted with concrete stimuli that trigger affective response." This is precisely what one would expect if the bias-construct were consciously inaccessible. Here, we see philosophical conceptual analysis leading to tangible empirical prediction.

However, if the above associative-propositional evaluation model of implicit bias is correct, none of this would serve as evidence of a consciously inaccessible state that has attributable representational content. This is because associative networks are not standardly regarded as representing the relations between concepts. This relates to a long-standing debate about whether implicit biases could be propositional or associative, with the general consensus among philosophers being that they must be propositional.[60] This discussion makes clear why this question matters. If the bias-construct is an associative network, then these cases would not serve as evidence for unconscious representational content attributable to the individual harboring the bias. Rather, they would be no different from the unconscious, "built-in" biases of the visual perceptual system.[61] If instead, the bias-construct is propositional, then in light of subjects' inability to report that generalizing content, it seems like we would have evidence of an unconscious mental state with attributable representational content. These biases would be bonafide cases of unconscious content.

Regardless of where one falls on this debate, though, the insights from the debate about unconscious bias highlight another reason why the question of content attribution within a bias's operation is important. Recall that one important impetus for attributing content was that reference to the relevant distal kinds were essential to explanations of how the capacity was operating. It's important that the representational contents, involving the relevant kinds, are the contents we take them to be. In the case of social bias, there are many potential kinds that could be causing us to treat individuals in ways that are indicative of

[59] This is similar to how vision scientists come to understand what assumptions are encoded in the biases of the visual perceptual system by investigating instances of when those biases form (in)accurate perceptions.

[60] Prominent defenders of the idea that implicit biases must be explicitly represented propositions include De Houwer (2014), Mandelbaum (2015), Levy (2015), and Karlan (2021).

[61] Such built-in biases are crucially marked by an in-principle indeterminacy in content. For more on this point, see Burge (2010, 404)'s discussion of perceptual formation principles not having privileged forms. This point is likewise congenial to Madva (2017, 88)'s claim that the content of implicit biases is indeterminate and only interpretable if "understood holistically and relationally, as part of broader cognitive-bodily-social-environmental systems."

bias against the social group they belong to, but that wouldn't necessarily indicate bias against that social group per se. This is because many biases toward individuals might be best explained by reference to seemingly innocuous causal features that merely correlate with the socially sensitive features of interest. Imagine, for example, a social media platform hiring workers. To make this decision, they choose an epistemically reasonable feature like candidates' being frequent users of the platform in question. This reasonable-to-use feature might, however, be highly correlated with some socially sensitive feature we think is problematic were they to use it explicitly in decision-making. For example, it might be that no elderly individuals use the platform; only much younger people do. Thus, their hiring decisions will—from a pure, causal, behavioral perspective—look exactly like (or similar enough to) decisions made on the basis of age.[62] As the discussion about unconscious perception above illustrated, in order to figure out whether reference to particular distal kinds is necessary, we need to identify cases where the two come apart and, in cases of near-perfect correlation, this will involve identifying cases of misrepresentation of certain social groups. Again, it is ultimately within the purview of social psychology to choose precise experimental designs. But philosophical reflection suggests that minimally, such designs must adjudicate between when the relevant kinds in a bias's operation (at any stage of the input, output, or construct) involve representations of the relevant social groups as such.[63]

[62] This is one reason why I object to accounts of bias that reduce it to mere behavioral dispositions (see, for example, De Houwer 2019). The reference to the appropriate representational contents is critical to the bias's being what it is.

[63] Munton (2021) presents an extremely compelling view of prejudice that is congenial to the functional account of bias that I offer. In both cases, the explanatory locus comes not in any particular mental attitude constituting the phenomenon in question (beliefs, affect), but rather in how attitudes are manipulated. For me, these manipulations will be realized in systematic transitions between mental states; for Munton, they are realized by salience structures that organize information as more or less readily accessible. For both, the explanatory import of these structures (functional/salience) come from their ability to facilitate non-inductive inference by making it computationally tractable (Munton 2021, 13; Johnson 2020c, 1195) and move to a higher level of abstraction that allows for multiple realizability (Munton 2021, 12, Johnson 2020c, 1215). However, it is on this point about content that her and my views most diverge. While we're both focused on mental processes as opposed to states and neither of us want to reduce these processes to stand-alone states, both views will depend in part on the mental states that these processes operate over. And whereas Munton at this point prescinds from details about the contents of those mental states, I believe they are central. (See Johnson 2020c, 1222-23 for a defense of maintaining representational inputs and outputs.) Reasons for this can be brought out by using Munton (2021, 7-8)'s example of Mark scrolling through PhilPapers and selecting only those articles written by men. This can occur, according to Munton, without Mark "even processing [the information that the papers are written by men and not women]" (p. 8) and despite the fact that "Mark does not form beliefs about the papers with female authors." (p. 7) Still, she claims, Mark would have a prejudice against women. I think the question of whether Mark has a social bias against women depends in part on whether he represents them *as women* (or that he represent the male authors *as men*). Ultimately, Munton's description of the case never clarifies what's going on in

In the case of the aforementioned experiments, that these kinds are the relevant kinds will be evidenced by subjects exhibiting bias toward not only members who belong to that social group, but also those who are mistakenly thought to belong to that social group. This is evidence that their bias aims to pick out members of a particular social group and, on the basis of some generalizing mechanism that systematically treats members of a that social group differently from non-members, treats those people differently. As in the case of unconscious perception, mere causal explanations to individuals won't do, since in these cases as we've described them, many of these individuals won't actually belong to the social group in question. So, like the illusion of a red berry, only explanations to what the states aimed to (but in this case failed to) pick out will do the trick. Representational content just is how we describe the objects and properties that some state aims to pick out. Thus, in those cases, content attribution will be necessary and, therefore, warranted.

5 Conclusion

In this paper, I have outlined two debates about unconscious content and the parallel problems that plague them both. I have then extended philosophical insights from one to shed light on possible paths for progress in the other. Throughout, the discussion has highlighted how these debates serve as paradigm case studies for investigating the fruitful exchange that can exist between philosophy and empirical science.

Mark's head when he displays the research patterns that he does, and the reading of her case is consistent with his picking out the male-authored papers because he represents the authors as male. However, the way she describes the case is consistent with another scenario (though admittedly unlikely) where Mark never even registers the names or genders of authors and, instead, he attends only to paper citation counts. If low-citation count is correlated with the author being a woman (as discriminatory practices might suggest), then while he's ignoring papers authored by women, it is not arguably because they are written by women. In such a case, Mark does not obviously have a bias against women. Or, at least, his bias is importantly different from cases of psychological bias wherein he does use representations of gender to sort papers. While I'm sure that in cases where discrimination against women ultimately explains why the properties an individual uses picks out the individuals that it does, we would all want to say that something problematic is occurring (see my discussion of "the proxy problem" in Johnson 2020a and Johnson 2022). Still, in a case where Mark doesn't even know the genders of the authors whose papers he's reading, at least some would reasonably hesitate to say that it is his prejudice against women that explains his behavior. Thus, minimally, I believe Munton (2021)'s account must include in its salience structures representational contents (however minimal) that represent the individuals as members of the relevant social group in the states that psychological operations (of saliency) operate over. Though, admittedly these questions warrant more discussion than I can provide here. I thank Jessie Munton for many insightful conversations about the role (or lack thereof) of representational content in psychological explanations of bias and prejudice.

6 Bibliography

Antony, L. (2001). Quine as Feminist: The Radical Import of Naturalized Epistemology. In Antony, L. and Witt, C. E., editors, *A Mind Of One's Own: Feminist Essays on Reason and Objectivity*, pages 110–153. Westview Press.

Antony, L. (2016). Bias: Friend or Foe? In Brownstein, M. and Saul, J., editors, *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, pages 157–190. Oxford University Press.

Banaji, M. R. and Greenwald, A. G. (2000). Pride and prejudice.

Banaji, M. R. and Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press, New York.

Beeghly, E. (2015). What is a Stereotype? What is Stereotyping? *Hypatia*, 30(4):675–691.

Berger, J. (2020). Implicit attitudes and awareness. *Synthese*, 197(3):1291–1312.

Berger, J. (2022). Kinds of Consciousness. In Young, B. D. and Dicey Jennings, C., editors, *Mind, Cognition, and Neuroscience: A Philosophical Introduction*, pages 251–266. Routledge.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(02):227.

Block, N. (2007). Overflow, access, and attention. *Behavioral and Brain Sciences*, 30(5-6):530–548.

Block, N. (2015). The Anna Karenina Principle and Skepticism about Unconscious Perception. *Philosophy and Phenomenological Research*, pages n/a–n/a.

Bodenhausen, G. V. and Gawronski, B. (2013). Attitude Change. In Reisberg, D., editor, *The Oxford Handbook of Cognitive Psychology*, pages 957–969. Oxford University Press.

Buehler, D. (2014). *Psychological Agency: Guidance of Visual Attention*. Doctoral Thesis, University of California, Los Angeles, Los Angeles, CA.

Burge, T. (2010). *Origins of objectivity*. Oxford University Press, Oxford.

Burge, T. (2020). Entitlement: The Basis for Empirical Warrant. In Graham, P. J. and Pedersen, editors, *Epistemic Entitlement*, pages 37–142. Oxford University Press.

Burge, T. (2022). *Perception: first form of mind*. Oxford University Press, Oxford, first edition edition. OCLC: on1227381102.

Cameron, C. D., Brown-Iannuzzi, J. L., and Payne, B. K. (2012). Sequential Priming Measures of Implicit Social Cognition: A Meta-Analysis of Associations With Behavior and Explicit Attitudes. *Personality and Social Psychology Review*, 16(4):330–350.

Carruthers, P. (2017). Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures? *Review of Philosophy and Psychology*, 9(1):51–72.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation: Implicit evaluation. *Social and Personality Psychology Compass*, 8(7):342–353.

De Houwer, J. (2019). Implicit Bias Is Behavior: A Functional-Cognitive Perspective on Implicit Bias. *Perspectives on Psychological Science*, 14(5):835–840.

Dennett, D. C. (1969). *Content and Consciousness*. Routledge and Kegan Paul.

Dunton, B. C. and Fazio, R. H. (1997). An individual difference measure of motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, 23:316–326.

Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. MIT Press.

Gawronski, B. (2019). Six Lessons for a Cogent Science of Implicit Bias and Its Criticism. *Perspectives on Psychological Science*, 14(4):574–595.

Gawronski, B. and Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5):692–731.

Gawronski, B. and Bodenhausen, G. V. (2014). Implicit and Explicit Evaluation: A Brief Review of the Associative-Propositional Evaluation Model: APE Model. *Social and Personality Psychology Compass*, 8(8):448–462.

Gawronski, B., Hofmann, W., and Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, 15(3):485–499.

Gawronski, B. and Payne, K. (2010). *Handbook of Implicit Social Cognition: Measurement, theory, and applications*. Guilford Press.

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27.

Hahn, A. and Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes? *Behavioral and Brain Sciences*, 37:28–29.

Hahn, A. and Gawronski, B. (2019). Facing One's Implicit Biases: From Awareness to Acknowledgment. *Journal of Personality and Social Psychology*, 115(5):769–794.

Hahn, A. and Goedderz, A. (2020). Trait-unconsciousness, State-unconsciousness, Preconsciousness, and Social Miscalibration in the Context of Implicit Evaluations. *Social Cognition*, page 19.

Hahn, A., Judd, C. M., Hirsh, H. K., and Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3):1369–1392.

Hillard, A. L., Ryan, C. S., and Gervais, S. J. (2013). Reactions to the implicit association test as an educational tool: A mixed methods study. *Social Psychology of Education*, 16(3):495–516.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., and Schmitt, M. (2005). A Meta-Analysis on the Correlation Between the Implicit Association Test and Explicit Self-Report Measures. *Personality and Social Psychology Bulletin*, 31(10):1369–1385.

Howell, J. L., Collisson, B., Crysel, L., Garrido, C. O., Newell, S. M., Cottrell, C. A., Smith, C. T., and Shepperd, J. A. (2013). Managing the Threat of Impending Implicit Attitude Feedback. *Social Psychological and Personality Science*, 4(6):714–720.

Howell, J. L., Gaither, S. E., and Ratliff, K. A. (2015). Caught in the Middle: Defensive Responses to IAT Feedback Among Whites, Blacks, and Biracial Black/Whites. *Social Psychological and Personality Science*, 6(4):373–381.

Howell, J. L. and Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *British Journal of Social Psychology*, 56(1):125–145.

Irvine, E. (2013). *Consciousness as a Scientific Concept: A Philosophy of Science Perspective*. Springer Netherlands, Dordrecht.

Jiang, Y., Costello, P., and He, S. (2007). Processing of Invisible Stimuli: Advantage of Upright Faces and Recognizable Words in Overcoming Interocular Suppression. *Psychological Science*, 18(4):349–355.

Johnson, G. (2022). Proxies Aren't Intentional; They're Intentional. *Unpublished Manuscript*.

Johnson, G. M. (2020a). Algorithmic bias: on the implicit biases of social technology. *Synthese*.

Johnson, G. M. (2020b). Are Algorithms Value-free? Feminist Theoretical Virtues in Machine Learning. *Unpublished Manuscript*.

Johnson, G. M. (2020c). The Structure of Bias. *Mind*, 129(516):1193–1236.

Karlan, B. (2021). The Rational Dynamics of Implicit Thought. *Australasian Journal of Philosophy*, pages 1–15.

Kentridge, R. W. (2015). What is it like to have type-2 blindsight? Drawing inferences from residual function in type-1 blindsight. *Consciousness and Cognition*, 32:41–44.

Klotz, W. and Neumann, O. (1999). Motor Activation Without Conscious Discrimination in Metacontrast Masking. *Journal of Experimental Psychology: Human Perception and Performance*, 25(4):976–992.

Krickel, B. (2018). Are the states underlying implicit biases unconscious? – A Neo-Freudian answer. *Philosophical Psychology*, 31(7):1007–1026.

Krickel, B. (2022). The Unconscious Mind Worry: A Mechanistic-Explanatory Strategy. *Philosophy of Science*.

Kriegel, U. (2009). *Subjective Consciousness: A Self-Representational Theory*. Oxford University Press.

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous*, 49(4):800–823.

Madva, A. (2017). Social Psychology, Phenomenology, & the Indeterminate Content of Unreflective Racial Bias. In Lee, E. S., editor, *Race as Phenomena: Between Phenomenology and Philosophy of Race*, pages 87–106. Rowman & Littlefield International.

Mandelbaum, E. (2015). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous*, 50(3):1–30.

Marshall, J. C. and Halligan, P. W. (1988). Blindsight and insight in visuo-spatial neglect. *Nature*.

Monteith, M. J., Voils, C. I., and Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, 19(4):395–417.

Munton, J. (2021). Prejudice as the misattribution of salience. *Analytic Philosophy*, page phib.12250.

Nier, J. A. (2005). How Dissociated Are Implicit and Explicit Racial Attitudes? A Bogus Pipeline Approach. *Group Processes & Intergroup Relations*, 8(1):39–52.

Norman, L., Akins, K., Heywood, C., and Kentridge, R. (2014). Color Constancy for an Unseen Surface. *Current Biology*, 24(23):2822–2826.

Peters, M. A. K., Kentridge, R. W., Phillips, I., and Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, 2017(1).

Phillips, I. (2015). Consciousness and Criterion: On Block's Case for Unconscious Seeing. *Philosophy and Phenomenological Research*, DOI:10.1111/phpr.12224.

Phillips, I. (2018). Unconscious Perception Reconsidered. *Analytic Philosophy*, page phib.12135.

Phillips, I. (2021). Blindsight is qualitatively degraded conscious vision. *Psychological Review*, 128(3):558–584.

Phillips, I. and Block, N. (2017). Debate on Unconscious Perception. In Nanay, B., editor, *Current Controversies in Philosophy of Perception*, pages 165–192. Routledge, New York, NY.

Prinz, J. J. (2015). Unconscious Perception. In Matthen, M., editor, *The Oxford Handbook of Philosophy of Perception*, pages 371–392. Oxford University Press, Oxford.

Quilty-Dunn, J. (2019). Unconscious perception and phenomenal coherence. *Analysis*, 79(3):461–469.

Rivers, A. M. and Hahn, A. (2018). What Cognitive Mechanisms Do People Reflect on When They Predict IAT Scores? *Personality and Social Psychology Bulletin*, page 15.

Rosenthal, D. (2005). *Consciousness and the Mind*. Clarendon Press, Oxford.

Sellars, W. (1956). Empiricism and the Philosophy of Mind. In *Science, Perception and Reality*. Routledge and Kegan Paul.

Stich, S. P. (1978). Beliefs and subdoxastic states. *Philosophy of Science*, pages 499–518.