Implicit Cognition and Unconscious Mentality

Tim Crane and J. Robert Thompson

Abstract

Theorists commonly postulate unconscious mental states and processes but are unable to articulate what it means to be unconscious. We dispute the standard view of the relationship between conscious and unconscious mentality, and with it, the standard view of the relationship between consciousness and intentionality. The second is to lay out several options for replacing the standard view, ones that allow for substantive differences between conscious and unconscious mentality. The third is to sketch the foundations of a unifying conception of the unconscious across the various disciplines which study the mind, focusing on the nature of interpretation and representation. Along the way, we apply these conjectures to examples of implicit cognition.

1. The Unconscious

It is now a commonplace amongst those who study the mind that it is largely unconscious, with only a small part of it manifesting itself in our conscious lives. Cognitive scientists routinely postulate unconscious states and processes among the central psychological machinery. The idea that the perceptual system, for example, makes 'unconscious inferences' has been around since Helmholtz (1867), and is part of the orthodoxy in computational theories of vision (Marr 1982).

Cognitive psychologists standardly appeal to unconscious priming effects and subliminal perceptions (Kihlstrom 1987). The puzzling phenomenon of blindsight seems an example of perception which is unconscious in some sense (Weiskrantz 1986), and some theories of perception treat the faculty itself as essentially unconscious (Burge 2010). Unconscious implicit biases and heuristics which affect our behavior and judgement are among the best-known hypotheses of social psychology (Wilson 2002) and behavioral economics (Tversky and Kahneman 1974).

In theoretical linguistics, the program of generative grammar explains human linguistic behavior in terms of stored, unconscious knowledge of grammatical rules (Chomsky 1980). In the philosophy of mind, the central mental states such as belief or desire (what Bertrand Russell (1921) called the 'propositional attitudes') are generally treated as essentially unconscious states, characterized by their causal or functional role.

Perhaps more familiar in popular circles is the unconscious as conceived by psychoanalysis, essentially involving the repression of desires and memories, which affect our behavior but are difficult to bring to consciousness (Freud 1915). Many of these psychological phenomena contribute to the popular image of the mind as an iceberg, with only its tip visible in consciousness and all the real action going on underneath.

There appears to be a bewildering variety of phenomena which the study of the mind classifies as unconscious, but does anything unite all these phenomena? Does the unconscious have an essence? Can there be a general theoretical account of unconscious mentality?

We proceed in this chapter with three aims. The first is to dispute the standard view of the relationship between conscious and unconscious mentality, and with it, the standard view of the relationship between consciousness and intentionality. The second is to lay out several options for replacing the standard view, ones that allow for substantive differences between conscious and unconscious mentality. The third is to sketch the foundations of a unifying conception of the unconscious across the various disciplines which study the mind. Along the way, we apply these conjectures to examples of implicit cognition.

2. The Very Idea of Unconscious Mentality

To get an adequate theoretical overview of what the unconscious means in these contexts, it helps to have a brief account of how the various conscious/unconscious distinctions in philosophy and psychology arose and developed from the late 19th century to the present day (for a longer account, see Crane (2020)). Without this context, it is difficult to locate where accounts of the unconscious go astray.

2.1 A short history

Psychology as a science began in the late 19th century chiefly as the study of conscious mental phenomena, but the emphasis was shifted to the study of behavior alone during the behaviorist period (roughly from the 1920s until the late 1950s). However, the return of mentalism or cognitivism in psychology after the Second World War did not reinstate the study of conscious phenomena as its chief concern. Instead, the psychology that emerged largely involved the attribution of systems of mental representation to subjects or to their brains to explain subjects' behavior — systems which are mostly unconscious.

This focus on representation appeared in philosophy as well, treating intentionality — what Franz Brentano called 'the mind's direction upon its objects' — rather than consciousness as its central focus. The object of a mental state, or its 'intentional object', is what it concerns, or is about, or is otherwise directed on. The intentional content, on this view, is the way the object is represented in the intentional state — the same object can be represented in different ways, and different objects can be represented in the same way. Many philosophers defended the idea that intentionality is what is distinctive of all the things we classify as mental (see Chalmers 2004; Crane 2003, 2009). Though the notion of unconscious

mentality had been around in philosophy in some form or another at least since Leibniz (1704), drawing a sharp distinction between intentionality and consciousness allowed 20th century philosophers a straightforward way of accounting for unconscious mentality — consciousness was not essentially intentional, and intentionality was not essentially conscious.

One result of drawing this distinction between consciousness and intentionality is that consciousness came to be conceptualized in predominantly sensory terms. Paradigmatic conscious states were bodily sensations like pains, and visual and other sensory experiences, selected for this role because they appeared to be characterizable in terms unrelated to any intentional or representational content they might possess. So such 'qualitative' or 'phenomenal' states are thought by many to be, for all intents and purposes, conscious states. A corollary of drawing this distinction so sharply is that because conscious intentional states were hard to incorporate within this framework, intentional states were treated as, for all intents and purposes, unconscious.

Drawing this sharp distinction enabled theorists to distinguish between conscious and unconscious mentality. But, such accounts resulted in a notion of conscious mentality that was ready-made for sensory states while remaining largely silent both about forms of nonsensory consciousness as well as what unconscious mentality was like. Even as unconscious mentality was invoked in an expanding number of contexts, it was modeled upon conscious mentality and treated as though it was not truly distinctive. As one pioneer in exploring unconscious mentality puts it:

> "This 'nothing special' line of argument is a direct entailment of taking the stance that consciousness is primary and that the default position should be

> that conscious processes lie at the heart of human cognition. From this perspective, unconscious, implicit functions are dealt with derivatively and virtually all interesting cognitive functions are to be seen as dependent on conscious processes" (Reber 1993: 25).

This *primacy of the conscious* did not deny the existence of the unconscious, *per se*, but it quietly shaped conceptions of the unconscious, treating conscious cognition as the default framework through which *all* mentality was to be understood (even the unconscious).

2.2 Ramifications

Although Reber argues that the unconscious/implicit holds a more legitimate claim to primacy than the conscious/explicit, one need not take a stand on primacy to note that conscious thought appears to be a bit of a latecomer to the cognitive scene. Hence, it is at least worth considering whether unconscious thought is able to achieve what it can because it operates in a manner lacking the particularities imposed by conscious thought. As an evolutionarily older mode of thinking with a wider scope than what thought happens to appear in consciousness, it may share few operational features with the conscious realm.

It is important, then, to make explicit the often-unarticulated assumptions involved in the primacy of the conscious. Even if humans are, for obvious reasons, more familiar with the conscious realm, the primacy of the conscious needs to be justified. It is not inevitable to begin with conscious mentality and treat the unconscious states and processes as though they were more of the same, but just lacking some qualitative 'glow' or 'buzz'. In fact, conscious mentality serves as a

poor guide to the nature of the unconscious, given the range of paradigm unconscious states and processes.

Models of consciousness were developed in a way to account for sensations. So, what kind of guide can they be for characterizing aspects of our mentality that aren't sensations? If we are right that theorists who believe in the primacy of the conscious tend to draw a sharp distinction between consciousness and intentionality, then they also tend to hold that intentional states are deemed to be primarily unconscious, perhaps essentially unconscious. But, if there is no obvious model of conscious intentional states to shape their characterization of their unconscious nature, how are we supposed to characterize unconscious mentality for this range of nonsensory states? This puzzle about conscious and unconscious mentality stems from attempts to shoehorn intentional states into a perspective that sees the conscious as primary but utilizes a conception of consciousness that is ill-suited for its explanatory aims.

Most current attempts to explain the unconscious are a result of a specific and implausible picture of consciousness that arose out of the behaviorist movement in philosophy and psychology (Crane 2020). It is critical to note that this picture is not mandatory. For one thing, these attempts fail to explain conscious intentionality. And without this conscious framework to adjust in order to explain the nature of the intentional, it's not clear how they can explain unconscious intentionality either. So, we need to rethink conscious intentionality, or we need to rethink unconscious intentionality — or both. Philosophy of mind in recent years has begun to concentrate on the former task (e.g., Bayne and Montague (2011), Farkas (2008), Kriegel (2013)), but in in this chapter, we concentrate on the latter.

2.3 Basic elements

With these origins in mind, we can place the account of unconscious mentality on a stronger footing. Let's begin by considering an incorrect, but instructive, view of the relationship between mentality and consciousness. John Searle (1992) has famously argued that consciousness is the only true mark of the mental; there are no unconscious mental states, there are only states of the brain which have the disposition to produce conscious states. Defenders of this position would need to demonstrate that the range of phenomena included in this volume are either not unconscious or not mental, which would be a tall order. But, although Searle's position is surely incorrect and is rejected by a majority of philosophers and psychologists, it presents a challenge to this majority to specify precisely what alternative they are defending. In particular, the idea of unconscious mentality requires clarification in two dimensions: what it means for an unconscious mental state or event or process to be *mental* and what it means for a mental state or event or process to be *unconscious*.

'Intentionalism', as we use the term here, is the view that all mental phenomena are intentional. Intentionalism gives a ready answer to the question of what makes unconscious states *mental* — they have intentional contents, or (in other words) they represent their objects in certain ways. Although we assume this intentionalist view in what follows, the conclusion of the chapter does not depend on it. Someone who rejects intentionalism could nonetheless accept the view of unconscious intentionality proposed here. They would have then to give a separate account of the mentality of the unconscious in non-intentional terms. Intentionalism holds that all mental states and processes have intentionality—both those that are conscious and those that are not. Hence, attempts to explain the conscious/

unconscious distinction solely in terms of the absence or presence of intentionality will not be viable. In what follows, we stress that the critical distinction to elucidate is that between conscious and unconscious intentionality.

3. Unconscious Intentionality

Among those who have studied unconscious intentionality, two broad approaches have emerged (cf., Katsafanas 2016). What we shall call the 'Dominant Approach' is uninformative and deflationary: unconscious mentality is just mentality that lacks consciousness. This is clearly in line with assumptions about the primacy of the conscious. The unconscious functions more or less like the conscious mind, simply without the presence of consciousness. The other approach denies this mentality-minus-consciousness claim. The unconscious functions according to its own special governing principles. In this section, we will give reasons in favor of the alternative approach that unseats the primacy of the conscious.

According to the Dominant Approach, unconscious representational states share their essential representational nature with their conscious counterparts; they just lack whatever it is that makes conscious states conscious. For philosophers who think of consciousness in terms of 'qualia', this will mean that unconscious states have intentionality but lack qualia; for higher-order thought theories, unconscious states will be ones which are not the objects of higher-order thought (Carruthers 2003, 2011; Rosenthal 2005); for theorists who postulate an ambiguity in 'consciousness' (Block 1995) states may be 'phenomenally' conscious and not 'access' conscious, or access conscious and not phenomenally conscious.

The notion of consciousness presumed here is (for the most part) modeled on sensations, so the Dominant Approach seems best poised to explain what an

unconscious sensation would be—it is a sensory state that lacks its conscious-making feature. This is, perhaps, what occurs in cases of blindsight described in §1. The Dominant Approach does little to characterize the nature of these unconscious intentional states, except to say what they lack, but this is still somewhat helpful in dealing with current mental episodes like sensations and perceptual experiences like blindsight. Nevertheless, an account of the unconscious also needs to deal with persisting or 'standing' states (like beliefs and intentions, for example).

Such standing states cannot be an afterthought for a theory of unconscious intentionality. Yet if we model consciousness on sensory states, this looks inevitable. The Dominant Approach must claim that an unconscious standing state can be understood as a conscious standing state that lacks consciousness. But what is a conscious standing state, i.e., in what sense can consciousness be present or absent for such states? Many, if not most, standing states *are* unconscious in their very nature (Crane 2013), so a proper account of unconscious intentionality needs to explain them.

Consider the paradigmatic example of a standing mental state—belief. Although it is sometimes said that people have conscious beliefs, this is a problematic idea. When it is said that a subject has conscious beliefs, what is typically being identified is not a belief or belief state, but an episodic judgment or assertion by the subject. That episode is justified or grounded by the agent's beliefs or is a report that came about because of her beliefs, but these are not conscious versions of beliefs, or beliefs with some conscious-making feature added. Unlike episodes of sensations, there are no episodes of believing, *per se*, conscious or not. This critical difference renders the Dominant Approach largely silent about the nature of unconscious beliefs. If there are not conscious beliefs, then one cannot

characterize the unconscious versions of these by taking a conscious version and removing its conscious-making feature. It's the features it possesses as a standing state that make it an unconscious belief, not the absence of some conscious-making features that a singular episodic occurrence of it might possess.

To be sure, it is somewhat cumbersome to distinguish between the standing belief states and an episode in which some judgment related to them is delivered, but this critical difference is precisely what is being muddled in the Dominant Approach, and what prevents the Dominant Approach from accounting for unconscious standing states. This stems from a commitment to the primacy of the conscious, where a model of *standing* belief states is erroneously shaped by a model of *episodic* judgment. According to the mentality-minus-consciousness approach, unconscious beliefs are explained as something like an episode of conscious judgment that has had its conscious element removed. An episodic sensory model of consciousness is thereby shaping our picture of unconscious belief in problematic ways.

Taking conscious mentality as a model for all mentality distorts the phenomena—elements are treated as episodic when they are not, as akin to sensory episodes or by postulating "a psychological structure which corresponds in a more or less direct way with the structure of conscious judgement or assertion" (Crane and Farkas 2022: 36-7). This seems innocent enough at first, in line with the primacy of the conscious, but as useful as these might be in commonplace explanations of behavior involving beliefs and desires, not all behavior can be attributed to episodes that parallel this doxastic/conative episode structure, but at the unconscious level. What's occurring at the unconscious level may not be episodic, and/or it may involve elements that diverge in significant ways from the doxastic/conative elements that

appear in our conscious level explanations. This primacy of the conscious is engrained enough either to escape notice or to feel unproblematic as the lens to view the unconscious, but it has profound impacts on how we view unconscious mentality.

There is a more helpful way to characterize the relationship between the conscious and the unconscious. Standing states are not, by default, conscious, and it is not clear how one would take a conscious standing state, remove some conscious-making feature, and have an unconscious standing state as a result. Nor would it seem that one can take the unconscious standing state and add some conscious-making feature and render it the conscious version of that state.

A pair of methodological decisions drive this distortion. The first is the assumption that the conscious and the unconscious realms operate using similar states and processes, that one can use a model of the conscious as a guide to the unconscious. The second specifies which similarities can be anticipated—that the features of conscious occurrent judgment will have corresponding features present among the unconscious intentional states. Yet, there are reasons to think the realms are distinct, and there are reasons to think that the occurrent judgment model is misleading.

There should be room for a radically different approach that takes seriously the hypothesis that the unconscious mind differs profoundly from the conscious in the way it represents the world, that unconscious mental representation works in very different ways from representation in consciousness. The unconscious is the basis of our psychological organization, but it may not be organized in the way that our conscious minds are.

Our discussion above focuses on the need for an account of unconscious intentionality to handle standing states. Such standing states play essential roles in our mental lives and our accounts of behavior, so they demand a proper treatment. But most of the mechanisms and states attributed by cognitive science are unconscious, whether they are occurrent or standing, so a proper account of the reality of these mechanisms and states requires a robust account of unconscious intentionality.

4. Nondeflationary Accounts

In what follows, we outline a range of nondeflationary alternatives that allow for the possibility that the unconscious bears little resemblance to the conscious. Our primary goal is to establish this approach as superior to those approaches closely aligned with the primacy of the conscious. With this alternative approach established, theorists can pursue a number of paths in exploring the different ways that the unconscious might be organized *differently* from the conscious.

4.1 Option One: A common alternative

The most common alternative to the Dominant Approach asserts that the unconscious is truly distinctive from the conscious but holds that the unconscious realm possesses states and processes that are characterized in contrast to those that exist in the conscious realms. For example, one can adopt a view of conscious mentality that involves rational, deliberative, and propositional thought while stipulating that the unconscious realm involves nonrational, nondeliberative, nonpropositional thought. The two realms operate according to different principles.

To better understand this approach, consider the phenomenon of implicit bias. A widely discussed case is that of implicit racist beliefs. It is often supposed that in such cases, a subject has consciously held egalitarian beliefs, but unconsciously held racist aliefs (Gendler 2008), attitudes, or biases. It stands to reason that the consciously held beliefs involve a psychological structure with a propositional content that drives egalitarian verbal reports, whereas the unconscious associations between racial categories and evaluative terms drive different nonverbal behaviors, like the amount of time it takes to react to a stimulus or sort objects.

A proponent of the Dominant Approach might explain the discordant behaviors as stemming from a conflict between two beliefs, an egalitarian conscious belief and a racist belief. The racist belief could have been a conscious belief, but it contingently lacks the conscious-making feature. So, on this view, there are not two realms, one rational/evidence-sensitive/propositional and one irrational/evidence-insensitive/associationistic. Since all beliefs operate similarly, any discordant behavior in cases of implicit bias is due to whether the conscious-making feature is present (or not) in one belief or the other.

Many writers have indicated the benefits and drawbacks to handling these sorts of cases along the lines of the Dominant Approach — in adapting one's conception of belief to handle such cases, in developing additional "in-between" approaches (see Schwitzgebel (2010), Brownstein and Saul (2016a, 2016b)) — or along the lines of the alternative described here involving associationistic structures. Given our commitments laid out above, the reader should be able to infer which of these we find more plausible than others: it should be clear that we don't find explanations that focus on the presence or absence of consciousness for some beliefs as being particularly illuminating, or even coherent. Moreover, we think the

alternative sketched above is overly limiting as a general approach to the unconscious.

Our alternative approach would start with the point argued above: that belief should not be treated as a conscious level phenomenon. If so, then it is preferable treat the egalitarian element as a conscious episode of judgment that is in conflict with racist unconscious associations, rather than a conscious belief that is in conflict with unconscious associations. Secondly, the Dominant Approach tends to assume a sharp delineation between the states and processes that are sensitive to evidence/rational/propositional and those that are not. And this is paired with the expectation that the states and processes that are sensitive to evidence/rational/propositional will be the conscious ones. Current research into implicit bias indicates that the distinctions are far less tidy than this, and that the unconscious states and processes involved in these cases are sensitive to evidence in ways that are not captured by the alternative. For example, at least some of the elements involved in implicit bias appear capable of modulation by rational argumentation and/or logical interventions (e.g., Sullivan-Bissett, this volume Ch. 8; Mandelbaum 2016).

We would like to draw a more general moral here: what happens in cases of discordant behavior like those described as “implicit bias” is not necessarily a struggle between two realms, one conscious, propositional, evidence-sensitive, and one that lacks all of those features. Indeed, it may be that both conflicting elements exist in the absence of conscious awareness.

Crane and Farkas (2022) also note that the roots of discordant behavior can be combined in many ways that the alternative seems to overlook. The roots might not at all doxastic (e.g., emotions, associations), somewhat doxastic (e.g., aliefs), or fully doxastic (e.g., beliefs). The point is that implicit bias and similar discordant

phenomena are unlikely to be explained only as the result of two realms operating according to different principles, one that is rational, deliberative, propositional, and conscious and one that is nonrational, nondeliberative, nonpropositional, and unconscious.

4.2. Option Two: Inferential integration

Not every distinctive feature of an unconscious mental state or process can be characterized as resulting from their operating in nonrational, nondeliberative, or nonpropositional ways. This is perhaps the most serious limitation with Option One: some unconscious mental states are distinctively different from consciously available states, but not in a way that precludes them from participating in inferences, being sensitive to rational argumentation, or possessing propositional contents. One way of characterizing this distinctiveness is to say they lack a certain level of inferential integration that limits their functional profiles and thereby limits their availability for conscious access and verbal report. Another way of describing this limitation of integration is in terms of *participation*: the relevant knowledge or structures don't appear to participate in many projects (Miller 1997) or are largely "harnessed to [a] single project" (Wright 1986: 227). In this sense, what is implicit will not be cognitively or inferentially integrated, will rarely or never be accessed by some other system for some other purpose, and will rarely (if ever) be subject to verbal report.

Many examples of implicit cognition exhibit this sort of unconscious profile. As noted in discussing Option One, some cases of implicit bias, upon examination, may involve unconscious mental states and processes that fail to be nonrational, nondeliberative, or nonpropositional in the requisite ways. As noted in §1, many of the representations postulated by cognitive scientists involve unconscious structures.

Option One and the Dominant Approach suggest only two possibilities in such cases: either these structures are nonrational, nondeliberative, and nonpropositional, or they are just like conscious representations, only (contingently) lacking the conscious-making feature. Option Two rejects this dichotomy. The unconscious structures can share features with conscious mentality but differ in terms of their inferential integration—they are limited in terms of their participation but are unlike the unconscious elements proposed in Option One. On this view, there is a range of unconscious mentality that seems to differ from conscious mentality in ways that suggest that the presence of mere associations could not be the full story. At least some of the behaviors are rich enough to posit a range of full-blown mental states to the subject that are unconscious. These do not get reported as frequently (if ever) and have a limited impact on behavior, but there is reason to deem them as propositional.

Option Two allows for thoughts with propositional contents and inferential connections, but their limited integration explains many of the features found implicit cognition. For example, cognition that is merely implicit is often used by an agent but not acknowledged by that agent when verbally prompted. And such implicit understanding often appears somewhat sporadically—it is exploited in some contexts but not in others, even where it would have been helpful. But, the fact that some understanding does not arise in conscious or verbally reported forms in every possible circumstance does require us to conclude that this understanding is flawed, limited, must be insensitive to evidence, or is the result of mere association, and the like. Option Two explains why the presence of cognitive machinery that lacks a certain level of inferential integration would manifest itself in these ways that theorists have identified as implicit. To fully distinguish these sort of cases from those

covered by Option One, theorists will need to do the difficult work of showing that the unconscious understanding driving some bit of behavior is too rich and sophisticated in the requisite ways to come from non-evidence sensitive, nonpropositional mental structures. And it needs to explain how it differs from the Dominant Approach by demonstrating how a lack of inferential integration accounts for the phenomena better than mere absence of the conscious-making feature.

Option One views the unconscious realm as exclusively irrational and associationistic (both in its processes and their contents), but this is too narrow a conception to account for all cases of unconscious mentality. Option Two shows that while some unconscious mentality may take the form suggested by Option One, there will be some cases where the features suggested by Option One will not suffice. Although we think there are several domains in cognitive science where Option Two is a viable approach to adopt, detailed work needs to be done in such cases to characterize the particular features exhibited by the unconscious intentional states and identify why the alternatives are inadequate (see Thompson 2014 for an example of what this might look like in developmental psychology).

4.3. Option Three: A unifying interpretative difference?

The discussion above suggests a certain heterogeneity in the unconscious realm that might warrant a pluralistic approach—perhaps the unconscious operates in ways that differ from the conscious in some instances, and ways that share some similarities with the conscious in others. In this section, we consider the extent to which there may be a more profound distinction between conscious and unconscious mentality that is reflected in and captured by our interpretational practices that deal with the unconscious realm. This may indicate something important about the states

and processes themselves, about the nature of the sorts of representations that are brought to bear in these situations.

Implicit cognition is like any theoretical construct in psychology — it is described by a process of interpretation. This need for interpretation applies both to the familiar kinds of unconscious mental states which we attribute to ourselves and to others in our everyday psychological thinking (thoughts, feelings, desires, intentions etc.), as well as to the unconscious mental states which cognitive science attributes to the brain or the subject. On this approach, although there clearly is unconscious mental representation, it is less determinate and explicit than conscious representation, and requires interpretation in a way that consciousness does not.

Drawing from Crane (2017), the alternative sketched here is that unconscious intentional states bear a different relationship to their own interpretation than conscious intentional states do. The central hypothesis is that the interpretation of unconscious mental states — whether by subjects attributing mental states to other human beings and to animals, by subjects themselves in self-attribution of belief, or by scientists attributing representational states to mechanisms in the brain — imposes an order on something which is much less ordered, explicit and determinate than the representations we find in consciousness. This is the common thread, which links all applications of the idea of the unconscious in philosophy and the various branches of psychology — even as philosophy and psychology have failed in their attempts to conceptualize the unconscious. Here we briefly consider both the attribution of intentionality to ourselves, and the attribution of representation in cognitive science.

What is it to attribute an intentional state to oneself? And how does one know what one's own intentional states are? This question has been intensely debated in

analytic philosophy under the (not entirely accurate) heading of ‘self-knowledge’ (Cassam 1994, 2014; Macdonald et al. 1998). How can one know what one thinks, or feels, or wants? This question poses a problem because the usual mechanisms of knowledge seem to have no application here — we do not seem to know our mental states by perception, testimony or inference. Or at least, there is not one of these models which works for all cases of self-knowledge. Saying that we know our mental states by ‘introspection’ seems to name the process rather than describe an actual mechanism.

What does the process of the acquisition of knowledge of our unconscious states tell us about the nature of the states themselves? Standard approaches assume that the process of finding out what you think is a matter of finding out what is (so to speak) ‘already there’. There are relatively fixed facts about your dispositions and these facts line up in a straightforward (if complex way) with attributions of intentional states in the language of commonsense psychology. The only question is to figure out what these facts are. But finding out what you think seems to be a different thing from making up your mind. Compare practical reasoning. When you are figuring out what to do, you will often deliberate and weigh up the various options and arrive at a decision or the formation of an intention. Similarly, when you are figuring out what to believe, you weigh up the evidence and come to an opinion. But this is supposed to be a different process from finding out what you already believe, though it will normally draw on things you already believe.

The standard approaches assume, in short, that there is a sharp distinction between finding out what you think (and want etc.) and making up your mind. This assumption is questioned by Crane (2017), who argues not that there is *no* distinction at all between finding out what you think and making up your mind — that

would be absurd — but that it is not a sharp distinction. In other words, there will be cases where the affirmation of something in consciousness could be conceived of as making a judgement about something on which you are genuinely unclear, or it could be a matter of reporting a straightforward belief — and there need be no fact of the matter about which of these is the correct description (Moran 2001).

This suggests an account of self-knowledge that allows for self-knowledge to be the product of *self-interpretation*, where self-interpretation is an essentially creative enterprise. Interpretation shapes one's vague or inchoate unconscious mental reality into the more determinate, specific form of a conscious judgement. Consciously putting one's thoughts into words is perhaps the clearest example of this, but there are many others.

Recognizing the centrality of interpretation in the understanding of the unconscious also sheds light on the question of how to understand the representational content of the states that cognitive science attributes to the brain or its mechanisms. What does it mean to say that the brain or the visual system represents something in the world outside? A traditional view (defended by Fodor 1975, 1987) is that for psychological theories to be explanatorily useful, they must be literally true, and this implies that there must be distinct representational states within the subject. This theory was opposed by those who thought that explanatory usefulness does not imply that there are representations in this literal, concrete sense (see Dennett 1975). In an older debate, this latter view is sometimes called 'instrumentalism', as opposed to Fodor's 'realism'.

This old dispute between realism and instrumental has stagnated in recent years. Each position seems too extreme. Instrumentalism seems to commit too little on the underlying structure of mental states, while realism over-commits (though see

Quilty-Dunn and Mandelbaum 2018). This led some philosophers to create a middle path between the two extremes (for an early attempt, see e.g., Peacocke 1983).

Important work on the role of representation in cognitive science has been done more recently (e.g. by Shea 2018 and Rescorla 2020). But nonetheless it is fair to say that there is no general consensus about how to think about 'realism' about representation in cognitive science. This suggests that a new approach is needed. Since most of the mechanisms and states attributed by cognitive science are unconscious, a proper account of the reality of these mechanisms and states must draw on a general account of unconscious mentality.

Any approach to the role of representation in cognitive science should take on board developments in the philosophy of science. We believe that for this reason we should apply the idea of a model as used in the philosophy of science (Weisberg 2007) to the question of representation. A scientific model is a simplified, idealized description or other object which aims to identify and isolate features of the system under investigation, and to explain the system's behavior by looking at the behavior of the model. Thus Rutherford's solar system model of the atom was an idealized representation of the atom's structure, but one which enabled understanding, explanation and prediction.

Similarly, computational models of the mind or the brain are also simplifying descriptions of the activity of mental faculties, which enable understanding, explanation and prediction of that activity. The models attribute propositional contents to states of the brain in the way analogous to the way that measurements of physical magnitudes employ numbers (Matthews 2010). This suggests that there will not be a unique content to any given state: the precise content attributed will be

relative to the model. What is important is to show that this does not make the intentionality of the brain or cognitive system ‘unreal’ or ‘merely instrumental’.

On this picture of the place of representation in cognitive science, abstract propositions do play a role, in modelling the unconscious mental states and processes. But what they are modelling is in itself representational — and the models are used to isolate some aspect of that underlying representational structure. The case for saying that this is representational is based on the familiar fact that we cannot identify the ultimate task that the brain or organism is performing without talking in representational terms: e.g., the visual system’s ultimate role is to create a representation of the visible world (Marr 1982; Burge 2010). But different theorists will interpret this representational structure in different ways, and use different models, some of which will be better than others in understanding the various subsidiary tasks performed by the system.

In its broad outlines, this conception of the role of content in cognitive science owes a lot to Cummins (1989), Dennett (1981) and especially Egan (2012). But these ideas have often been misleadingly associated with ‘instrumentalism’ where that label carries the insinuation that the conception does not treat intentionality as sufficiently real. This is partly because the standard for an intentional state being ‘real’ has been set in an implausibly simplistic way by the Fodor-style realist, or by general metaphysical maxims associating reality with causal efficacy. What has gone wrong here is not the modelling picture, but the conception of what realism requires. We need a better picture of what it is for a representational state of the brain to be real, to understand the reality modelled by cognitive science: the representational reality of the mind is made more explicit and determinate by the theorist’s interpretation, which assigns specific contents to specific states in its model.

Finally, it is worth pointing out that this picture of the attribution of intentional states to ourselves and to others has a connection with certain central ideas from psychoanalysis. Psychoanalysis remains controversial of course, both as a therapy and as a theory — and we do not endorse any specific psychoanalytic theories. However, our approach shares with psychoanalysis the idea that the unconscious mind operates according to its own principles; mental representation works in a different way in the unconscious from the way it does in the conscious mind (cf., Katsafanas 2016). The relevance of psychoanalysis is that it puts at the heart of its theory and practice the fact that there are reasons why people do things which they do not themselves fully understand; and drawing out what these reasons are may involve imposing an order on something that does not in itself have such an explicit order. This is where conscious interpretation imposes an order on the relatively unformed and chaotic unconscious.

Despite the controversies surrounding psychoanalysis, many of the phenomena it attempts to explain are real, and it is an advantage of an account of the unconscious that it can make room for them. Our understanding of the role of interpretation, we claim, has the potential to provide a psychologically realistic account of the relationship between the various manifestations of the unconscious, while also preserving the distinctive unity of the unconscious mind and its distinction from consciousness.

5. Conclusion

Cognitive science needs a conception of unconscious mentality that serves its explanatory needs. To achieve this, theorists need to reconsider the primacy of the conscious and its Dominant Approach to unconscious mentality. Although we think

there is much to be gained by pursuing the approach sketched in §4.3, once theorists are able to view the unconscious realm in its own right, more adequate attention can be given to the different ways in which intentional states might be unconscious.

Related Topics: *Due to the centrality of consciousness to the study of implicit cognition, there are no principled grounds for excluding any of the other chapters in the Handbook.*

References

Bayne, T. and Montague, M., eds., 2011. *Cognitive phenomenology*. Oxford: Oxford University Press.

Block, N. 1995. "On a confusion about a function of consciousness". *Behavioral and Brain Sciences*, *18*: 227–247.

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

Carruthers P. 2003. *Phenomenal consciousness: a naturalistic theory*. Cambridge: Cambridge University Press.

Carruthers, P. 2011. *The opacity of mind: an integrative theory of self-knowledge*. Oxford: Oxford University Press.

Cassam, Q., ed., 1994. *Self-knowledge*. Oxford: Oxford University Press.

Cassam, Q. 2014. *Self-knowledge for humans*. Oxford: Oxford University Press.

Chalmers, D. 2004. "The representational character of experience". In B. Leiter, ed., *The future for philosophy*. Oxford: Oxford University Press: 153–181.

Chomsky, N. 1980. *Rules and representations*. New York: Columbia University Press.

Crane, T. 2003. "The intentional structure of consciousness". In A. Jokic and Q. Smith, eds., *Consciousness: new philosophical perspectives*. Oxford: Oxford University Press: 33–56.

Crane, T. 2013. "Unconscious Belief and Conscious Thought." In U. Kriegel, ed., *Phenomenal intentionality: new essays*. Oxford: Oxford University Press:156-73.

Crane, T. 2020. "A short history of the philosophy of consciousness in the Twentieth Century". In A. Kind, ed., *Philosophy of mind in the Twentieth and Twenty-First Centuries: the history of the philosophy of mind, vol. 6*. London: Routledge.

Crane, T. and Farkas, K. 2022. “The limits of the doxastic”. In U. Kriegel, ed., *Oxford studies in philosophy of mind,* vol. 2. Oxford: Oxford University Press: 36-57.

Cummins, R. 1989. *Meaning and mental representation.* Cambridge, MA: MIT Press.

Dennett, D.C.1975. “Brain writing and mind reading” In K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science,* 7: 403-15.

Dennett, D.C. 1981. “A cure for the common code”. In *Brainstorms.* Hassocks: Harvester: 90-108.

Egan, F. 2012. “Metaphysics and computational cognitive science: let's not let the tail wag the dog”. *The Journal of Cognitive Science, 13*: 39-49.

Fodor, J. A. 1975. *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. A. 1987. *Psychosemantics: the problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Freud, S. 1915. *The unconscious*. London: Hogarth.

Gendler, T. S. 2008. “Alief in action (and reaction)”. *Mind & Language*, *23*: 552–585.

Helmholtz, H. 1867. *Handbuch der physiologischen optik*. vol. 3. Leipzig: Voss.

Katsafanas, P. 2016. *The Nietzschean self: moral psychology, agency, and the unconscious*. Oxford: Oxford University Press.

Kihlstrom, J. F. 1987. “The cognitive unconscious”. *Science*, *237*: 1445–1452.

Leibniz, G.W. 1704, *New essays on human understanding*. Trans: P. Remnant and J. Bennett. 1981. Cambridge: Cambridge University Press.

Macdonald, C., Smith, B. C., and Wright, C. J. G., eds., 1998. *Knowing our own minds: essays in self-knowledge*. Oxford: Oxford University Press.

Mandelbaum, E. 2016. “Attitude, inference, association: on the propositional structure of implicit bias”. *Noûs*: 629–658.

Marr, D. 1982. *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco: W. H. Freeman and Company.

Matthews, R. 2010. *The measure of mind*. Oxford: Oxford University Press.

Miller, A. 1997. “Tacit knowledge”. In B. Hale and C. Wright, eds., *A companion to the philosophy of language*. Oxford: Blackwell.

Moran, R. 2001. *Authority and estrangement: an essay on self-knowledge*. Princeton: Princeton University Press.

Peacocke, C. 1983. *Sense and content: experience, thought and their relations*. Oxford: Oxford University Press.

Quilty-Dunn, J., and Mandelbaum, E. 2018. “Against dispositionalism: belief in cognitive science”. *Philosophical Studies*, *175*: 2353–2372.

Reber, A. S. 1993. *Implicit learning and tacit knowledge: an essay on the cognitive unconscious*. Oxford: Oxford University Press.

Rescorla, M. 2020. “Reifying representations”. In J. Smorthchkova, T. Schlicht, and K. Dolega, eds., *What are mental representations?* Oxford: Oxford University Press: 135-177.

Rosenthal, D. M. 2005. *Consciousness and mind*. Oxford: Oxford University Press.

Russell, B. 1921. *The analysis of mind.* London: George Allen and Unwin.

Schwitzgebel, E. 2010. “Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief”. *Pacific Philosophical Quarterly*, *91*: 531–553.

Searle, J. 1992. *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Shea, N. 2018. *Representation in cognitive science*. Oxford: Oxford University Press.

Thompson, J.R. 2014, “Signature limits in mindreading systems”. *Cognitive Science, 38:* 1432-1455.

Tversky, A. and Kahneman, D. 1974. “Judgment under uncertainty: heuristics and biases”. *Science*, *185*: 1124–1131.

Weisberg, M. 2007. “Who is a modeler?” *British Journal for the Philosophy of Science, 58*: 207-233.

Weiskrantz, L. 1986. *Blindsight: a case study and implications*. Oxford: Clarendon Press.

Wilson, T. D. 2002. *Strangers to ourselves: discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Wright, C. 1986. “Theories of meaning and speakers’ knowledge”. In *Realism, meaning, and truth*. Oxford: Blackwell.