**"Functionalist Theories of Consciousness"**

revised July 2007

to appear in T. Bayne, A. Cleeremans, and P. Wilken, eds., *Oxford Companion to Consciousness*

Bernard W. Kobes
Department of Philosophy
Arizona State University
Box 874102
Tempe, Arizona 85287-4102

kobes@asu.edu

The term *functionalism* occurs with distinct meanings in several disciplines; this article concerns functionalism as a philosophical theory, or better a family of theories, about the nature of mental states. At the core of functionalism is a metaphysics of mental types (§1, below). Perspectival and phenomenal aspects of consciousness pose vivid challenges for this account, but some broad functionalist strategies may be deployed in response (§2). Recently prominent functionalist theories of consciousness may be seen as implementations of these strategies (§3). Functionalism persists as a controversial backdrop for current themes in philosophy of consciousness (§4).

**§1. Functionalism as a metaphysics of mental types.**

Since the 1960s functionalism has been the most prominent of a set of *isms* (dualism, behaviorism, physicalism, ...) that address the mind-body problem. Consider the "metaphysical" question: in virtue of what are distinct particular mental states or events grouped together as instances of the same mental type? Functionalism denies that a given type of mental state is to be characterized in terms of its *intrinsic* properties, its constitution or ontology. Instead, mental types are fundamentally *relational*, where the relevant relata include input stimuli, output behaviors, and other mental states. The core idea is that a type of mental state is a type of causal role in a larger network linking inputs, outputs, and other internal states.

Computing theory supplied an early model for functionalist ideas in

philosophy of mind. The internal states of a Turing machine are defined by their role in the machine table, which defines all relevant internal states simultaneously, without vicious circularity, and in abstraction from their material realization. By analogy, *pain* can be characterized simultaneously with other internal states (e.g., pain-typical desires, emotional responses such as anxiety) in a causal network linking internal states to each other and to typical inputs (e.g., tissue damage) and outputs (e.g., avoidance behaviors), in abstraction from its material realization.

Contrast functionalism with two other two broadly materialist programs: behaviorism, and a physicalism that identifies mental types with neurological types. Functionalism corrects behaviorism's reluctance to postulate "inner" states, yet it articulates a level of description at which mental types and neurological types cross-classify. A human being, a non-human animal, a futuristic automaton, and an extra-terrestrial organism might all be in the inner mental state-type *pain* even if they share no relevant neurological state-type. In brief, mental state-types are *multiply realizable*. It is compatible with this, and many functionalists went on to assert, that each particular instance of pain— each pain *token*— is strictly identical with some neurological (or silicon, etc.) state or event token in the relevant human being, non-human animal, automaton, or extra-terrestrial. In that case we would have identity between mental tokens and neurological (or silicon, etc.) tokens, but no identity, or even universal correspondence, between mental types and neurological types.

The functional role that defines a mental type must be articulated by way of a prerequisite larger theory of the causal relations into which that mental type enters. One sort of functionalist theory— *psychofunctionalism*— sees the prerequisite causal theory as given by empirical cognitive science. Think of empirical sciences as revealing the hidden natures of the kinds they treat; the nature of water, for example, is empirically revealed to be $H_2O$. Mental kinds are empirically revealed by cognitive science to be functional in nature, according to psychofunctionalism.

Another sort of functionalist theory sees the required theory of causal relations as given by folk psychology, that network of causal relations our knowledge of which constitutes shared commonsense understanding of the mental, and which we presuppose in our practical explanations of human action. Or the causal theory may be restricted to just those causal relations entailed by the very meanings of the relevant mentalistic terms; these causal principles will be analytic— truths of meaning. *Folk-psychological functionalism* and *analytic functionalism*, as these latter two variants are termed, are best understood as aiming to directly explicate our mentalistic concepts, rather than the natures of mental kinds as such.

**§2. Functionalist strategies in response to the challenge of consciousness.**

Functionalism has always been motivated primarily by third-person, public aspects of mentality, but it has faced a series of vivid challenges keyed to first-person, perspectival or phenomenal aspects of consciousness. In F. Jackson's thought experiment, for example, Mary is brought up in a controlled environment in which she can have no chromatic color experience. Black and white visual experiences are adequate, however, for her training as a vision scientist, and indeed Mary learns all of the physical and functional facts about color experience. One day she escapes her controlled environment, looks at a ripe tomato, and learns what it is like to see red. Since she learns a new fact about color vision, after already knowing all the physical and functional facts, it seems to follow that some facts about color vision— what-it-is-like facts— are neither physical nor functional.

D. Chalmers, building on ideas from S. Kripke, argues that we can coherently conceive of a philosophical "zombie"— a creature molecule-for-molecule identical to an ordinary human being, but in which there is no phenomenal consciousness. Your zombie twin would perfectly replicate your physical and functional organization, down to the minutest details of input stimulation, output behavior, and neuro-functional intermediaries, but there is nothing that it would be like to be your zombie twin. Zombies are metaphysically possible— they might have existed, had the natural order been different. Chalmers concludes that no physical or functional organization can be metaphysically sufficient for phenomenal consciousness.

From the conceivability of zombies, J. Levine draws an epistemic rather than a metaphysical conclusion: we cannot explain, on the basis of the underlying physical or functional facts, the fact of a conscious state having a particular phenomenal character, rather than having a different phenomenal character or none at all. This *explanatory gap* is a permanent feature of our epistemic condition. The gap persists even if we postulate metaphysically necessary connections between the underlying functional facts and the facts of phenomenal consciousness, for we cannot explain why *those* connections hold and not others.

These thought experiments — and others, such as the *inverted spectrum*, and J. Searle's *Chinese Room* — put pressure on functionalists to attend to the phenomenal and first-person perspectival aspects of consciousness. In this effort the distinction between *phenomenal properties* and *phenomenal concepts* has come to loom large. Concepts are ways of thinking about entities or properties; the distinct concepts *water* and $H_2O$, for example, pick out the same entity or property in the world. Jackson's protagonist Mary has first, we may suppose, a functional concept of what it is like to see red; on exiting the controlled environment she gains

a phenomenal concept of, a new way of thinking about, what it is like to see red. But these two concepts pick out the same phenomenal property. Mary's visual experience of the ripe tomato enters into her phenomenal concept of what it is like to see red; she thus acquires fresh and vivid knowledge of what it is like. But the phenomenal *fact* that she thereby knows is the very fact that she previously knew via functional concepts. By insisting on a dualism of concepts rather than a dualism of properties or facts, a functionalist can interpret the case of Mary as consistent with the thesis that all properties and facts of consciousness are functional in nature.

Similarly, many functionalists have responded to the zombie thought experiment by arguing that the scenario is not conceivable in any sense strong enough to entail metaphysical possibility. Zombies are conceivable in a weaker sense that does not entail metaphysical possibility because we can refer to a single phenomenal property via distinct concepts. Imagining a zombie, we use functional concepts when specifying its construction, then use phenomenal concepts when denying it consciousness. Thus the dualism of concepts makes zombies weakly conceivable. But if a third-person functional concept and a corresponding first-person phenomenal concept pick out the same property of consciousness with respect to every possible world, then zombies are metaphysically impossible. The distinction between *first-person (subjective)* and *third-person (objective)* in philosophy of consciousness is fundamentally a distinction between kinds of concepts, not kinds of properties or facts, according to functionalism.

If some discomfort or sense of mystery still attaches to the idea that a phenomenal / perspectival fact could simply *be* a functional fact, that may be dispelled by conceptual or empirical progress in articulating the two sides of the identity. Two broad theoretical strategies dominate recent functionalist accounts of consciousness; these strategies are often deployed in tandem, with details varying from one philosopher to another. The first strategy, focused mainly on the phenomenal / perspectival side of the equation, is to *divide and conquer*; that is, to make appropriate distinctions between kinds of consciousness, or aspects of consciousness, and then to attack them piecemeal. Perhaps functionalism is true for some aspects of consciousness but not others.

The second strategy, focused mainly on the functional side of the equation, is to take *intentionality* as more basic than consciousness, and to explain consciousness in terms of intentionality. Intentionality is the "aboutness" of mental states, their property of referring to some thing, or representing some state of affairs. The thing referred to may or may not actually exist, and the state of affairs represented may or may not actually obtain. Hence intentional mental states may be

accurate or illusory, true or false, satisfied or unsatisfied. Suppose that intentionality can be explained functionally; causal theories of reference, or functional role theories of intentional content, may be recruited toward this end. Suppose further that consciousness can be explained in terms of intentionality. Conjoining these two suppositions yields a functionalist explanation of consciousness.

**§3. Implementing the strategies: theories and variations.**

As an example of the first strategy, consider N. Block's notorious distinction between phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness). A mental state is P-conscious in virtue of its experiential properties, broadly construed to include not only properties of sensations, feelings, and perceptions, but also those of thoughts, wants, and emotions. A mental state is A-conscious in virtue of being poised for direct control of reasoning and for rational control of action or speech. P-conscious states give rise to the so-called "hard problem" of consciousness, but are candidates for neurophysiological reduction. A-conscious states, by contrast, essentially make representations accessible to a larger system, and so may be understood in terms of functional role.

Block takes the P-vs.-A distinction to pick out not two aspects but rather two kinds of consciousness, or two senses of the word 'conscious'. (Chalmers draws a similar distinction.) It is conceptually possible, given Block's account, for a creature to have A-conscious states without ever having any P-conscious states. Yet it has seemed to many philosophers that, given Block's broad construal of the phrase *experiential property*, a creature with no P-consciousness during its entire career would not be conscious in *any* reasonable sense. This suggests that the P-vs.-A distinction is badly drawn. As a divide-and-conquer strategy, the P-vs.-A distinction meets further resistance from functionalist programs that aim to explain the phenomenal in terms of the intentional.

D. Dennett favors broadly functionalist metaphors for consciousness such as *cerebral celebrity*, *fame in the brain*, or *competition for clout.* Mental contents instantiated in the brain form coalitions and compete for control of action and verbal report. Coalitions that succeed in this competition count as having conscious contents. Those contents need not be transmitted to any part of the brain that implements consciousness; they need only be organized into successful coalitions. Such a functional process may be realized in humans by a kind of reverberation or amplification loop. Efficacy in feedback and control of action is key. As S. Hurley and A. Noë have argued, there need be no discrete input and output layers, and the

reverberation or loop may be conceived as extending out into the world.

Dennett's account is more than usually metaphorical. Yet it is functionalist in spirit, for it characterizes consciousness in terms of a *causal role*, hence at a more abstract level than that of a physiological *role-player*, and the possibility of physiologically diverse realizers is left open. By contrast, an account that simply identified consciousness with, say, synchrony of neural oscillation plus ventral stream activation (for visual contents), or "re-entrant" neural circuits, would not be functionalist. On the other hand, "global workspace" or "global broadcast" theories in psychology, and Schacter's Conscious Awareness System, are functionalist in spirit, in the manner of cerebral celebrity.

Philosophical critics worry that a zombie, lacking consciousness, could nevertheless instantiate such functional patterns. It is therefore unclear how cerebral celebrity or its ilk could simply *be* consciousness, as opposed to something normally accompanying consciousness though in principle separable. Moreover, examples of habituated stimuli, and perhaps also subliminal stimuli, suggest the possibility of phenomenality— the most philosophically puzzling aspect of consciousness— without cerebral celebrity. Dennett's discussions of phenomenal and perspectival aspects of consciousness are multi-faceted and pose interpretive difficulties. In some respects Dennett's account is neo-behaviorist and appears to treat phenomenal and perspectival aspects as less than robustly real.

One striking way to develop the second strategy mentioned above, that of explaining consciousness in terms of intentionality, takes phenomenal properties to be represented properties of ordinary external, non-mental things. Many contemporary philosophers, including G. Harman, F. Dretske, and M. Tye, accept some form of the *transparency thesis*, which insists that conscious experience never has any directly introspectible phenomenal properties of its own. What are often taken to be phenomenal properties of conscious experience are actually properties that experience represents external non-mental things as having. A conscious visual experience of a ripe tomato has no introspectible property of phenomenal redness; it merely represents, accurately or inaccurately, the tomato as red. Further, the represented property of redness may be identical to a complex physical property of external surfaces, a matter of reflectance frequencies. In the modern tradition we inherited from Descartes, phenomenal properties were "kicked upstairs" into the mind to make the non-mental world safe for mechanistic physics. The transparency thesis kicks *represented* phenomenal properties back out into the non-mental world, and lets them be identical to mechanistic physical properties.

The transparency thesis is a natural ally of *first-order representationism*, the

doctrine that the phenomenal character of consciousness is exhausted by— supervenes on and depends on— certain of its first-order representational properties. A *first-order* representational property is one that purports to represent external, non-mental objects or states. (A representational property that purported to represent a mental state or event would be termed *higher-order*.) But not all first-order representational properties contribute to phenomenal character. Which do, and which do not? According to M. Tye's first-order representationist theory of consciousness, phenomenal character is *poised*, *abstract*, *non-conceptual intentional content* (PANIC). Of these functional conditions, two are most relevant here: *poised* content stands ready to directly impact general cognition and action, assuming attention is properly focused and certain concepts are possessed; *non-conceptual* content has features for which the subject need not possess matching concepts. Thus on Tye's account the phenomenal character of consciousness is entirely a matter of properties of external, non-mental objects or states as represented in a certain functionally specific way.

Critics of Tye's theory cite the inverted spectrum, which seems to show the possibility of two subjects who are representationally exactly alike but who differ phenomenally. In a variant, *Inverted Earth*, the environment is imagined to vary so as to suggest that a subject on Earth and her twin on Inverted Earth may be phenomenally exactly alike but representationally dissimilar. Both inversion arguments exploit the alleged dependence of the purely phenomenal on the subject's central nervous system, and the dependence of representation on the history of the subject's environment. Discussion has also focused on whether the phenomenal character of sensations such as headaches and orgasms are truly representational, given that a conscious state's having a biological function does not yet amount to its being representational. Similar doubts arise with respect to the phenomenal character of emotions and moods.

*Higher-order representationism* also aims to explain consciousness in terms of intentionality, but unlike its first-order counterpart it takes mental states to be conscious in virtue of their being the intentional objects of other mental states. D. Armstrong and W. Lycan argue that a mental state *m* is conscious in virtue of *m*'s being the object of an inner perception or scanning process— a *higher-order perception* (HOP). Precursors of this idea can be found in the history of philosophy, most notably in Locke. Non-conscious perceptual representation, as for example in blindsight, is taken to be more fundamental than consciousness, and amenable to a functionalist treatment. Often perceptual representation is directed upon mundane external things, but there is no reason why it cannot also be directed instead upon

other mental states, in which case the represented mental states are conscious.

A variant of this idea is D. Rosenthal's *higher-order thought* (HOT) theory, which takes the relevant higher-order state to be an occurrent thought rather than a perception. A first-order mental state *m* is "like something" for the subject when *m* is the intentional object of a second-order thought. A second-order state may, in some cases, be the intentional object of a third-order state, but this occurs when the subject engages in conscious introspection, as a philosopher, psychologist, or poet might do, and not in everyday consciousness. Introspective consciousness reveals that the second-order state introduces no distinctive phenomenal quality of its own, not already present in the first-order state. Hence the need, according to Rosenthal, to make the higher-order states *thoughts* rather than perceptions.

Recall that according to functionalism, a mental token belongs to its mental type in virtue of the token's relational properties. In HOP and HOT theories the relevant type is simply that of a state's being *conscious*, which is not any intrinsic property but a matter of its being an intentional object of a distinct higher-order state. Note that the higher-order state itself need not be conscious, so there is no need for an infinite hierarchy of representations.

HOP and HOT theories both distinguish a state's having *phenomenal quality* from its being *phenomenally conscious*. The latter is explained in terms of higher-order representation, while the former is explained in some other physicalist or functionalist manner. The distinction entails the intelligibility of *unconscious phenomenal quality*, a striking and theoretically fruitful result, according to the theory's proponents, but a conceptual vulnerability according to some critics. Critics also argue that higher-order representationism, especially the HOT theory, is caught on the horns of a dilemma: either it over-intellectualizes consciousness in babies and lower animals such as mice or fish, or it implausibly denies them consciousness. Finally, a hallmark of intentionality is the possibility of error and illusion, so erroneous or illusory HOPs or HOTs must be intelligible and presumably actual. Again, proponents find this a striking and theoretically fruitful result, to be interpreted in light of cognitive science. Critics, however, charge confusion— an erroneous or illusory HOP or HOT would determine what it is like for the subject, but it is the first-order state, not the HOP or HOT, that is supposed to be conscious.

In response to these and other alleged difficulties, a variety of *self-representationist* or *same-order representationist* approaches are also being explored. R. Van Gulick's *higher-order global states* (HOGS) model, for example, may perhaps be classified as self-representationist. The central idea is that a mental state *m* becomes conscious in virtue of *m*'s being subject to a rich set of implicit

sub-personal processes that (a) recruit *m* into a globally integrated complex, and (b) amount to reflexive awareness of *m* itself. Condition (a) is a form of cerebral celebrity, condition (b) a form of representationism.

**§4 Further problems and prospects.**

The ferment over representationism shows the resilience of functionalist ideas in the face of *prima facie* damning thought experiments. One of the above theories may well emerge as rendering it plausible that certain kinds of phenomenal or perspectival facts simply *are* causal-role facts. Moreover, as of this writing (2007), functionalism persists as a backdrop to a range of current questions and obsessions in the philosophy of consciousness. Three categories may be briefly noted.

First, can the functionalist metaphysics be modified to accommodate a recognition that conscious worldly agency is typically temporally extended and involves dynamic sensory feedback? The "vehicles" of such conscious events are naturally seen as looping out into the world; the inputs and outputs of the functionalist metaphysics seem implausibly discrete buffers.

Second, can functionalism be coherently articulated with respect to the *first-person phenomenologies* of attention, of mental agency, of immediate self-knowledge, or of the (apparent, alleged) attribution in conscious color perception of "revealed", primitive, Edenic qualities to external things?

Third, richer integration with neuroscience is both desirable and inevitable. Functionalists distinguish between *core* and *total realizations* of conscious state *m*: the core is a physical state *n* that plays the distinctive causal role, and the total is a wider state within which *n* plays that role, sufficient for *n* to constitute an *m* token. What in the neurobiology of consciousness answers to such a distinction? And—Block's "harder problem" of consciousness— what basis could we ever have for attributing phenomenality to, or withholding it from, conceivable creatures that are functionally isomorphic to us but that have nothing like our nervous systems?

**References:**

Armstrong, David M. (1980), "What is Consciousness?", in Armstrong's *The Nature of Mind*, University of Queensland Press (pp. 55-67).

Block, Ned (2007), *Consciousness, Function, and Representation: Collected Papers, Volume 1*, MIT Press.

Burge, Tyler (1997), "Two Kinds of Consciousness", in Ned Block, Owen Flanagan, and Güven Güzeldere, eds., *The Nature of Consciousness: Philosophical Debates*, MIT Press (pp. 427-433). Reprinted in Burge, *Foundations of Mind: Philosophical Essay, Volume 2*, MIT Press (2007), pp. 383-391.

Chalmers, David J. (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press.

Dennett, Daniel C. (2005), *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*, MIT Press.

Gennaro, Rocco J., ed. (2004), *Higher-Order Theories of Consciousness: An Anthology*, John Benjamins Publishing Company.

Harman, Gilbert (1990), "The Intrinsic Quality of Experience", in James E. Tomberlin, ed., *Philosophical Perspectives 4* (pp. 31-52).

Kriegel, Uriah and Kenneth Williford, eds. (2006), *Self-Representational Approaches to Consciousness*, MIT Press.

Lycan, William G. (1996), *Consciousness and Experience*, MIT Press.

Papineau, David (2002), *Thinking About Consciousness*, Oxford University Press.

Rosenthal, David M. (2005), *Consciousness and Mind*, Oxford University Press.

Shoemaker, Sidney (2003), "Content, Character and Color", in E. Sosa and E. Villanueva, eds., *Philosophical Issues, 13, Philosophy of Mind*, pp. 253-278.

Tye, Michael (2000), *Consciousness, Color, and Content*, MIT Press.

Van Gulick, Robert (2006), "Mirror, Mirror — Is That All?", in U. Kriegel and K. Williford, eds., *Self-Representational Approaches to Consciousness*, MIT Press (pp. 11-39).