Philosophy Compass 1/1 (2006): 58–64, 10.1111/j.1747-9991.2006.00008.x

Consciousness, Theories of

Uriah Kriegel
University of Arizona/University of Sydney

Abstract

Phenomenal consciousness is the property mental states, events, and processes have when, and only when, there is something it is like for their subject to undergo them, or be in them. What it is like to have a conscious experience is customarily referred to as the experience's *phenomenal character*. Theories of consciousness attempt to account for this phenomenal character. This article surveys the currently prominent theories, paying special attention to the various attempts to explain a state's phenomenal character in terms of its representational content.

Phenomenal consciousness is the property mental states, events, and processes have when, and only when, there is something it is like for their subject to undergo them, or be in them. There is something it is like to smell coffee brewing. Having the experience of smelling coffee brewing is thus a phenomenally conscious state. What it is like to have a conscious experience is customarily referred to as the experience's *phenomenal character*. Theories of consciousness attempt to account for this phenomenal character.

Such theories are often divided, in the first instance, into *physicalist* and *anti-physicalist*. Physicalist theories attempt to account for phenomenal consciousness in (micro)physical terms. Anti-physicalist theories claim that this is impossible. Perhaps a more fundamental division is into *reductive* and *non-reductive* theories. Reductive theories attempt to account for phenomenal character in non-phenomenal terms. Non-reductive ones do not. Whether the relevant phenomenal terms can in turn be accounted for in (micro)physical terms is left an open question at this stage.

Although most philosophical work has concentrated on reductive theories, some has been devoted to non-reductive ones as well. By claiming that phenomenal consciousness cannot be reductively accounted for in non-phenomenal terms, a non-reductive theory treats phenomenal consciousness as a *fundamental* feature of the world. It thus considers that there are (at least) two fundamental kinds of feature in the world: physical and phenomenal; for this reason, this sort of view is known as *dualism* (from the Latin word for "two").

As just noted, fundamental features cannot be accounted for in terms of other features. Thus if phenomenal consciousness is fundamental, it cannot possibly be accounted for in non-phenomenal terms. But this does not mean

we cannot have a theory of it. One may theorize about phenomenal consciousness by specifying the laws of nature that govern the causal (or other) interactions of phenomenal events (i) among themselves and (ii) with non-phenomenal events. Such a specification would constitute an account of phenomenal consciousness without attempting to reduce consciousness to something non-phenomenal (Chalmers 1996, 2002).

Reductive theories attempt to identify an ostensibly non-phenomenal feature of mental states and account for consciousness in terms of it. Most then proceed to account for the relevant non-phenomenal feature in purely (micro)physical terms; the result is known as *Physicalism*. Traditionally, two kinds of feature have been appealed to by reductivists: *functional* and *representational*. It has also been traditionally assumed that the relevant functional and representational features are amenable to physicalist treatment.

Mental states have typical causes and effects. A feeling of sadness, for instance, might be caused by emotional injury and cause going on a shopping spree. A mental state's *functional role* is given by a subset of all the state's typical causes and effects. *Functionalism* is the theory that attempts to account for the phenomenal character of conscious states in terms of their functional role. (See Dennett 1981, 1991; Baars 1988, 1997; it is also possible to interpret the "enactive" theory of consciousness – as in Noë 2004 – as a sophisticated version of functionalism.)

Mental states typically also represent all sorts of things. A visual experience of a rainbow, for example, represents the rainbow, its colors, shape, etc. A mental state's *representational content* is what it represents or purports to represent. *Representationalism* (sometimes known as "First-Order Representationalism," for reasons that will become evident momentarily) is the theory that attempts to account for a conscious state's phenomenal character in terms of its representational content (Dretske 1995; Tye 1995, 2000).

According to *Higher-Order Representationalism*, mental states are not conscious in virtue of *representing*, but in virtue of *being represented*. In other words, they are conscious because they are themselves the representational contents of higher-order representations. For example, a conscious feeling of a tickle behind one's ear involves one's awareness of one's tickling feeling, and that awareness is a matter of the tickling feeling being targeted by a higher-order representation (Armstrong 1968; Rosenthal 1986, 1990, 2002; Lycan 1996; Carruthers 2000; Van Gulick 2001, 2006).

Finally, according to the *Self-Representational Theory* (or "Self-Representationalism"), whatever else a conscious experience represents, it always also represents itself; and it is in virtue of thus representing itself that it is conscious. On this view, then, all and only conscious states are *self-representing*. For example, when you sit in the tub and consciously stare at your toes, it is true both that you are aware of the toes, and that you are aware (though more dimly) of your visual experience of the toes. According to the self-representationalist, this is because your conscious experience at the time represents both your toes and itself (Smith 1986, 1989; Kriegel 2003).

I will now turn to a very summary survey of the merits and demerits of each of these theories. I will present what I take to be the strongest line of argument in favor, and then the strongest line against, each of the theories just sketched. For all these arguments and counter-arguments, there are innumerable objections, rejoinders, comebacks, modifications, and complications that have been explored in the literature, but which we will not have occasion to discuss here.

The main argument *for dualism* is the *argument from the conceivability of zombies*. We can readily imagine, in all seriousness, creatures that are physically indistinguishable from us but are not conscious, in the sense that they do not have conscious experiences. If such perfect "zombies" are indeed possible, it would mean that our consciousness is something "extra," something over and above all the physical facts about us. For we could have been physically exactly the same and yet have no consciousness. (See Chalmers 1996. For a different key argument, see Jackson 1984.)

The main argument *against dualism* is the *argument from causal efficacy*. The charge is that dualism entails the thesis that conscious states do not have the power to affect the physical domain. The thought is that the physical domain is "causally closed" – every physical event is fully caused by some physical event – and therefore, on the assumption that physical events are unlikely to be systematically fully caused by two independent sets of causes, non-physical events would normally be deprived of any causal efficacy vis-à-vis the physical domain. This consequence is however extremely unintuitive: it certainly seems that when I consciously decide to raise my arm, my conscious decision causes my arm's subsequent motion (Kim 1989, 2001).

The main argument *for physicalism* is the fact that science has managed over and again to account for initially mysterious and apparently recalcitrant phenomena in purely (micro)physical terms. It would be odd if consciousness stood out, all said and done, as the only phenomenon defying the trend (Smart 1959). The arguments *against physicalism* are basically the arguments *for* dualism.

We may call the main argument *for functionalism* "the argument from everything we always wanted." On the one hand, we want to believe that there are no non-physical phenomena in the world; on the other, we want to believe that consciousness is in some way independent of brute physical matter. If phenomenal character were just functional role, this might just be the case. Functional roles, being roles, must be occupied. It is possible to hold that, on the one hand, the functional role that defines phenomenal consciousness could in principle be occupied by any number of different physical features, so consciousness is independent of – is "something more" than – any one of them; but that, on the other hand, all the possible occupants of that functional role must be physical features, and therefore there are no non-physical features in the world (Putnam 1967).

The main argument *against functionalism* is that we can imagine cases in which (i) the right functional role is not accompanied by phenomenal character; or (ii) the same functional role is accompanied by different phenomenal characters. Thus, we can imagine (i) a gigantic nation whose citizens interact in a way that mimics the functional interaction of neurons in the brain, without that nation having conscious experiences (as a nation); and (ii) two persons whose color spectrum has been completely inverted, so that one experiences green when the other experiences red, but because the comparative relations among their respective experiences are exactly the same, their experiences have the exact same functional role (Block 1978, Shoemaker 1975).

The main argument *for representationalism* is the *argument from transparency*. Consider your visual experience of the computer in front of you. If you try to turn your attention away from the computer and onto your experience of it, you find that the only feature of your experience you can detect is its representational content: you detect that it is an experience *of a computer*. In other words, when you examine your experience of the world, you cannot but see the world right through it – as though the experience was in itself transparent. To suppose then that the phenomenal character of your experience is nonetheless distinct from its representational content is to suppose that we are under a massive illusion regarding the phenomenal character of our experiences. That seems implausible on the face of it (Harman 1990; Tye 2000).

As with functionalism, numerous arguments by counter-example have been offered *against representationalism* (Peacocke 1983; Block 1990). But they are all relatively tendentious. However, there is available a more principled argument. Presumably, any object or feature in the world can be represented either consciously or unconsciously. Yet representationalism, by its insistence that the difference between states that are phenomenally conscious and states that are not is in what they represent (their representational content), is committed to the existence of objects or features which only lend themselves to conscious representation. This is a most implausible commitment (Kriegel 2002; Chalmers 2004).

The main argument *for higher-order representationalism* starts from the observation that conscious states are states we are aware of having. The notion of a conscious state of which the subject is totally unaware seems almost like a contradiction in terms. Indeed, what makes a mental state conscious is that one is aware of it in the right way. For it is only when a mental state is represented by the subject that there is something it is like *for the subject* to be in that state. Now, being aware of something is a matter of having a representation of it (being aware of a tree involves having a representation of the tree). So the subject's awareness of her conscious state is a matter of her having a representation of that state. It follows that what makes a mental state conscious is the subject's having a higher-order representation of it (Lycan 2001; Rosenthal 2002).

The main argument *against higher-order representationalism* is the *argument from targetless higher-order representations*. Higher-order representations, like their first-order counterparts, can misrepresent. Moreover, they may not only misrepresent the *properties* of their targets, but also their very *existence*. It follows from higher-order representationalism that targetless higher-order representations result in a subjective impression of being in a conscious state without actually being in one. This means that, absurdly, there is no conscious state the subject is in, but there is something it is like for the subject at that moment. Thus, suppose a person has a higher-order representation to the effect that she is having a taste of white chocolate. But she is not in fact having a taste of white chocolate. She is not having a taste of anything. According to higher-order representationalism, this person does not have any conscious experience, but it seems to her as if she does. In other words, the person is not conscious, but is nonetheless under the impression that she is. This sounds absurd (Byrne 1997; Neander 1998; Levine 2001).

The main argument *for self-representationalism* has the form of a dilemma. Conscious states are states we are aware of, and therefore represent ourselves as having. But a conscious state can be represented either (i) by some higher-order representation; or else (ii) by itself. Since (i) leads to such problems as are manifest in the argument from targetless higher-order representations, and these are avoided by (ii), we should accept (ii). If a mental state is conscious in virtue of being represented by itself, it cannot be that the state is represented to exist when in fact it does not exist. For if it did not exist it would be unable to represent itself (or anything else for that matter). The upshot is that the only way to preserve the idea that conscious states are states we are aware of having without falling prey to the problems of targetless higher-order representations is to hold that conscious states are self-represented, hence self-representing (Kriegel 2003).

The main argument *against self-representationalism* concerns the very notion of self-representation, which is sometimes considered incoherent, unintelligible, or inconsistent with accepted doctrines about representation. Thus it seems essential to the notion of representation that there be a distinction between what is being represented and what does the representing, but no such distinction could apply in the case of self-representation. In the same vein, it seems that representation requires at least a minimal causal relation between the represented and the representing; but nothing could bear any causal relation to itself. For many, the main attraction of representational accounts of consciousness is that they pave the way for physicalism, since it is thought that representation will succumb to a purely physical account in mostly information-theoretic terms; but there is no clear yet non-trivial sense in which items can be said to carry information about themselves (Levine 2001, 2006).

The debate over the relative merits and demerits of the various theories of consciousness is not about to resolve itself anytime soon. Nor is it possible

to rule out the emergence of new contenders on the scene. As the debate intensifies, it is likely to become more and more technical, as well as make ever-increasing contact with empirical work in the cognitive sciences. Yet the persistent dissatisfaction in the philosophical community with extant theories of consciousness suggests that the debate is unlikely to proceed only on the technical level.

Bibliography

Armstrong, D. M. 1968. *A Materialist Theory of the Mind*. New York: Humanities Press.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.

Baars, B. 1997. *In the Theater of Consciousness: The Workspace of the Mind*. Oxford and New York: Oxford University Press.

Block, N. J. 1978. "Troubles with Functionalism," *Minnesota Studies in the Philosophy of Science*, 9, 261–325.

Block, N. J. 1990. "Inverted Earth," *Philosophical Perspective*, 4, 52–79.

Byrne, A. 1997. "Some Like It HOT: Consciousness and Higher Order Thoughts," *Philosophical Studies*, 86, 103–29.

Carruthers, P. 2000. *Phenomenal Consciousness*. Cambridge: Cambridge University Press.

Chalmers, D. J. 1996. *The Conscious Mind*. Oxford and New York: Oxford University Press.

Chalmers, D. J. 2002. "Consciousness and Its Place in Nature," in *Philosophy of Mind*, ed. D. J. Chalmers. Oxford and New York: Oxford University Press.

Chalmers, D. J. 2004. "The Representational Character of Experience," in *The Future for Philosophy*, ed. B. Leiter. Oxford: Oxford University Press.

Dennett, D. C. 1981. "Towards a Cognitive Theory of Consciousness," in *Brainstorms*. Brighton: Harvester.

Dennett, D. C. 1991. *Consciousness Explained*. Boston, MA: Little Brown.

Dretske, F. I. 1995. *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Harman, G. 1990. "The Intrinsic Quality of Experience," *Philosophical Perspectives*, 4, 31–52.

Jackson, F. C. 1984. "Epiphenomenal Qualia," *Philosophical Quarterly*, 34, 147–52.

Kim, J. 1989. "Mechanism, Purpose, and Explanatory Exclusion," *Philosophical Perspectives*, 3, 77–108.

Kim, J. 2001. "Lonely Souls: Causality and Substance Dualism," in *Soul, Body and Survival: Essays in the Metaphysics of Human Persons*, ed. K. Corcoran. Ithaca: Cornell University Press.

Kriegel, U. 2002. "PANIC Theory and the Prospects for a Representational Theory of Phenomenal Consciousness," *Philosophical Psychology*, 15, 55–64.

Kriegel, U. 2003. "Consciousness as Intransitive Self-Consciousness: Two Views and an Argument," *Canadian Journal of Philosophy*, 33, 103–32.

Levine, J. 2001. *Purple Haze: The Puzzle of Consciousness*. Oxford and New York: Oxford University Press.

Levine, J. 2006. "Conscious Awareness and (Self-)Representation," in *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

Lycan, W. G. 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.

Lycan, W. G. 2001. "A Simple Argument for a Higher-Order Representation Theory of Consciousness," *Analysis*, 61, 3–4.

Neander, K. 1998. "The Division of Phenomenal Labor: A Problem for Representational Theories of Consciousness," *Philosophical Perspectives*, 12, 411–34.

Noë, A. 2004. *Action in Perception*. Cambridge, MA: MIT Press.

Peacocke, C. A. B. 1983. *Sense and Content*. Oxford: Clarendon.

Putnam, H. 1967. "The Nature of Mental States." Originally published as "Psychological Predicates," in *Art, Mind, and Religion*, ed. W. H. Capitan and D. D. Merrill. Reprinted in The Nature of Mind, ed. D. M. Rosenthal. Oxford: Oxford University Press.

Rosenthal, D. M. 1986. "Two Concepts of Consciousness," *Philosophical Studies*, 94, 329–59.

Rosenthal, D. M. 1990. "A Theory of Consciousness," *ZiF Technical Report*, 40, Bielfield, Germany. Reprinted in 1997. *The Nature of Consciousness: Philosophical Debates*, ed. N. J. Block, O. Flanagan and G. Guzeldere. Cambridge, MA: MIT Press.
Rosenthal, D. M. 2002. "Explaining Consciousness," in *Philosophy of Mind*, ed. D. J. Chalmers. Oxford and New York: Oxford University Press.
Shoemaker, S. 1975. "Functionalism and Qualia," *Philosophical Studies*, 27, 291–315.
Smart, J. J. C. 1959. "Sensations and Brain Processes," *Philosophical Review*, 68, 141–56.
Smith, D. W. 1986. "The Structure of (Self-)Consciousness," *Topoi*, 5, 149–56.
Smith, D. W. 1989. *The Circle of Acquaintance*. Dordrecht: Kluwer Academic Publishers.
Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.
Tye, M. 2000. *Consciousness, Color, and Content*. Cambridge, MA: MIT Press.
Van Gulick, R. 2001. "Inward and Upward – Reflection, Introspection, and Self-Awareness," *Philosophical Topics*, 28, 275–305.
Van Gulick, R. 2006. "Mirror Mirror – Is that All?" in *Self-Representational Approaches to Consciousness*, ed. U. Kriegel and K. Williford. Cambridge, MA: MIT Press.

