



Subjectivity as Self-Simulation: Virtualising the Cartesian Theatre

Roly Perera^{1,2}(✉) 

¹ Department of Computer Science and Technology, University of Cambridge,
Cambridge, UK

roly.perera@cl.cam.ac.uk

² School of Computer Science, University of Bristol, Bristol, UK
roly.perera@bristol.ac.uk

Abstract. Scientific theories of subjective experience, such as such as Global Workspace theory and Attention Schema theory, are being co-opted as architectural proposals for artificial systems. A lack of consensus on what a scientific theory of subjectivity should actually explain will make it hard to evaluate any such artificial systems, however. In this paper, we adopt a naturalistic starting point, rejecting the idea that science must explain why we *have* subjective experiences, and instead suggest that science need only explain why we *take ourselves* to be having subjective experiences. But we resist the idea that this move leads to some kind of illusionism; instead we propose a more functionalist approach that saves certain key realist intuitions.

Specifically, we argue that an idea from computer science, namely *virtualisation*, and the associated distinction between abstract content and implementation mechanism, sheds light on two characteristic features of subjectivity: first, the fact that we seem to have *dense* (everywhere rich and contentful) subjective lives, in contrast to the sparser view emerging from cognitive neuroscience; and second, the fact that we take our subjective lives to be broadly *private*. To the extent that these are robust (functional) features of our self-theorising, they need to be accounted for by any naturalistic theory of subjectivity, and are likely to be important characteristics of any agents with artificial subjectivity.

1 Phenomenal Functionalism

Scientific theories that purport to explain subjectivity¹ in natural minds, such as Baars' Global Workspace theory [3] and Graziano's Attention Schema theory [16], are informing the next generation of architectures for artificial minds. Determining whether and to what extent these new systems might be conscious will require moving towards consensus on how subjectivity is to be understood as a natural phenomenon.

¹ As Metzinger [22] notes, "consciousness" is an overloaded term; here we will use the term "subjectivity" and will mean first-personal, phenomenal experience.

Starting with Ryle [29,30], various well-known perspectives on subjectivity, including Dennett's Multiple Drafts [12] and Metzinger's Phenomenal Self Model [23,24], involve a core naturalistic assumption. They either propose, or tacitly assume, that if we account for all the "heterophenomenological" data – all the public, in-principle-observable data, such as our dispositions to produce certain verbal reports – there would be nothing left to explain. The advantage of this methodological starting point is clear: no more than the usual resources of science are required to explain subjectivity, for example no additional meta-physical assumptions or axiomatic stipulations.

Scientifically well-motivated they may be, but these accounts remain largely unsatisfying to the phenomenal realist. Dennett's position attempts to undermine the (Cartesian) idea of subjectivity as a rich inner image; critics have characterised his view as "consciousness explained away" [8]. (Strong) illusionism in the style of Frankish [15] denies that there are any such things as "phenomenal feels"; according to critics, this denies the obvious, e.g. that we genuinely feel pain [9]. Metzinger's claim that "nobody has ever *been* or *had* a self" ([21], p. 1) is similarly challenging to the realist intuition.

In this paper we take the naturalistic starting point seriously, but reject the idea that the only alternative to phenomenal realism (at least as traditionally construed) is illusionism. Instead, we propose that the self-model approach of Metzinger, Blanke, and others [5,24], as well as the Multiple Drafts perspective of Dennett, are better framed in terms of *phenomenal functionalism*, a somewhat deflationary stance but one which retains broad compatibility with the contentful part of the realist intuition. On this view phenomenal feels are real, causal things with central functional roles in human behaviour.

Taking the naturalising move seriously means this. To explain all the phenomena in question, both objective *and* subjective, a science of subjectivity need only explain *why we take ourselves* to be having subjective lives in the way that we do. By treating our (perhaps sub-personally constituted) *beliefs about ourselves* as the data to be explained, this move still takes our intuitions about our subjective lives seriously, but makes them amenable to scientific explanation by rejecting the idea that we take them literally. Instead it characterises this rich collection of "self-suppositions" – the things we take to be true of ourselves, for example that, privately in my subjective world, things look or feel a certain way to me – as the data to be explained.

This move is a well-known trick in computer science, usually attributed to David Wheeler [20], of adding a level of indirection in order to treat something as data. Here, the indirection consists in taking our intuitive certainty *about* the explanandum, and treating that intuitive certainty as *part of* the explanandum, thereby reducing the problem to something within the reach of existing science. It works as a methodological move, we suggest, because it is not possible to distinguish an agent which "has" a subjective life, from one which merely *takes itself* to be having one; and crucially, this is true not only from the vantage point of any third party, but also from the vantage point of the agent itself. Thus a theory of how these (sophisticated and non-trivial) self-suppositions come about

would be sufficient to explain not only all our public self-reports, but also our ostensibly privately-held beliefs about ourselves as well.

Even the realist will concede that such an explanation would be sufficient if the goal were “merely” to explain how our first-personal lives *seem* to us. The residual disagreement with the (strong) illusionist would be ontological: while the realist insists that they *really are* in pain, the illusionist denies there is any such thing, despite conceding that it is “obvious” that it seems like there is [9]. But as there are no data, objective or subjective, that can distinguish these two possibilities, the illusionist appears to be in a more parsimonious position.

We propose a different tack. Rather than downgrading the nature of experience from “real” to “illusory”, we will argue that *taking oneself* (in a suitably rich way) to be having a subjective life is, in a scientifically meaningful sense, *constitutive of* having one. This is not just semantic wordplay: we will argue that there is real (albeit functionally characterised) thing in the world that is a subjective self, with a phenomenal life; a phenomenal life that is moreover (in a functional sense) a rich inner image that the agent is able to internally “observe” and reflect upon.

This is a less problematic response to the realist because it fully acknowledges, without rescinding the naturalistic move, the richness and apparent privacy of subjectivity. A lot, however, hinges on how we unpack “in a suitably rich way”. Calling the way we take the world to look to us, or how we feel, “suppositions” or “intuitions” is misleading to the extent that it presents such things as merely propositional; actually their content is much richer. According to our everyday experience, our subjective lives are rich, imagistic, egocentric *worlds* [22]. On this functionalist view, then, what needs explaining is not just *that* we take ourselves to be having subjective lives, but also how the subjective lives we take ourselves to be having come to have the specific character that they do.

One further point of clarification is needed before we proceed. A worry might be that, in order to *suppose* oneself to be having subjective experiences, one must already *have* subjective experiences; perhaps one can only “self-theorise” in this sort of sophisticated way if one is conscious in the first place. If the very concept we are trying explain is required to formulate the explanandum, then the prospects for a non-circular explanation start to look poor. For this particular problem to be avoided, it must be possible for the personal-level “beliefs” in question (those that, on the view presented here, are *constitutive of* the subjective life of the agent) to be implemented using sub-personal mechanisms; we must think of them as sub-personal inferences, rather than “beliefs” in the cognitive sense. (One might of course *also* hold similar beliefs cognitively, i.e. at the personal level; but those personal-level beliefs must be quite different in character and cannot be part of the explanation of how we come to be conscious.)

Incorporating the richness of our subjective lives into the explanandum helps naturalise a key part of the realist intuition. But at least two significant challenges remain. The first is how to reconcile the rather dense (continuous and highly contentful) nature of these subjective worlds with the (by comparison sparse) picture emerging from neuroscience and the psychology of perception.

We return to this in detail in Sect. 2, where we will attempt to resolve this using the idea of *virtualisation*, which allows for dense content to be delivered using sparse mechanisms. Whereas Dennett and others have used the evidence for sparsity to argue against the plausibility of imagistic first-personal content [13, 27], we will suggest that the perspective of virtualisation allows sparsity to be understood as an *implementation strategy* for dense content.

The second challenge is to accommodate the common realist view that some key aspects of consciousness have an essentially private, non-functional character. These appear to be widely shared intuitions, that by their very nature present a significant practical impediment to treating the *having* of such intuitions as data, as highlighted by Chalmers [9]. In Sect. 3 we will suggest that this points to a central architectural feature of systems with subjectivity: an organisational structure which embeds (a simulation of) the agent into a simulated world, a simulation which assigns to the agent a subjective life whose content is immediately present to the agent.

2 Virtualising the Cartesian Theatre

One way of restating the claim so far is that a science of subjectivity should reject the idea that we *observe ourselves* being conscious (those observations providing the data to explain). To save all the phenomena, it is sufficient to recognise that we *take ourselves to be observing ourselves* being conscious (and then to treat this self-conceptualisation of ourselves as “inner observers” as the data to explain).

If we accept this move, then the task facing a science of subjectivity is to explain how we come to take ourselves to be inhabiting an apparently private egocentric world, with Metzinger’s Self-Model theory [23, 24] being the canonical example of such an approach. In this section we unpack this obligation in terms of a characteristic feature of these subjective worlds: the fact that, for the most part, we take subjective experience to be “dense” (everywhere populated with rich content). This seems to be in tension with the picture emerging from cognitive neuroscience, which is much sparser. Whereas much of the scientific debate concerns whether consciousness is dense or sparse, here we suggest that another perspective from computer science, namely *virtualisation*, may offer a way to reconcile these two viewpoints.

2.1 Finding Out as Filling In: Phenomenal Content on Demand

The question of whether subjective content is sparse or dense has been much debated, with a growing body of evidence pointing to sparsity. Influential empirical work by Dehaene and others [10] on attentional blink and attentional blindness showed how phenomenal content is not globally reliable: sub-personal attentional mechanisms play a big role in determining subjective content in the presence of competing stimuli. Experiments based on *phi* (illusory movement) and neon colour spreading, discussed by Dennett, Kinsborne and others [13, 27],

showed that the brain more typically *finds out* (makes an inference about how the world looks or seems) than *fills in* (constructs a neural representation isomorphic to the content). Many cognitive scientists and philosophers have argued that these findings show that the (ostensibly) rich, continuous nature of visual experience is illusory.

But do these discoveries really mean that we are wrong about our own subjective lives? Here we suggest not: it is perfectly possible for the represented *content* to be dense, but for the mechanisms for delivering that content to be *sparse*. That is not to say that clever experiments cannot reveal the underlying sparsity; but what they reveal (on this view) are *implementation mechanisms* for dense abstractions, plus insights into the *operating norms* outside of which those abstractions become unreliable. This can be understood in terms of another key concept from computer science, namely the distinction between “abstraction” and “implementation”, and the related idea of *virtualisation*.

To illustrate this we will contrast *dense* and *sparse* representations of a 2D matrix. A *dense* representation stores every element of the matrix explicitly in a 2D data structure. Any element i, j of the matrix is retrieved by looking it up in the data structure using the i and j coordinates. A *sparse* representation, by contrast, is not a data structure. Rather, it is a *service* providing a *capability*. When supplied with i, j coordinates, it is able to deliver the element at i, j , but it may do so for example by utilising a representation which only stores the non-zero elements (with other compression strategies being possible). The full 2D content is *operationally*, or functionally, present, in the sense that an arbitrary query over that content can be satisfied, but there is no saturated (dense) representation being continually maintained. To a user of the matrix, the query interface is a functional abstraction that hides the distinction between sparse and dense implementations, at least under normal conditions; “finding out” here is a *kind of* filling in.

Here the sparse representation is in a sense a “virtualisation” of the 2D matrix: it relies on the fact that a matrix can be given a *functional specification* as a container of numbers, characterised by the ability to deliver the number associated with any chosen i, j coordinate.² Our suggestion is that we should think of subjective content as a similarly virtual space of content, with the organism participating on both sides of the abstraction: at the personal level (as a subjective agent), *consuming* the rich content, and at the sub-personal level, *producing* that content using sparse inferential mechanisms, such as saccades, attentional shift, and other sub-personal processes. This offers the prospect of resolving the tension between these two perspectives by allowing personal-level content to be dense and sub-personal delivery mechanisms to be sparse.

Indeed, on both theoretical and practical grounds, we would expect cognitive mechanisms to use an optimised mixture of sparse and dense representations: sparse representations when it would be too costly to do upfront work

² The “virtual” here is not in the sense of *virtual reality*, but rather in the sense of *virtual memory*, a service that provides the abstraction of a large, contiguous block of memory, even if physical memory is smaller or fragmented.

that might turn out to be unnecessary (such as interpolating missing information), and dense representations when the environment is predictable enough for some upfront work to pay off. And indeed there is considerable evidence from neuroscience that in some cases neural structures do interpolate missing information [31]. But the key point remains: to a *consumer* of the content, such as an introspective mechanism that queries the content of our subjective world, the particular implementation details are mostly unobservable.

The substantial line of evidence for sparsity, then, rather than revealing the sparsity of subjective content, reveals the sparsity of the inferential mechanisms that underlie that content, and the science of subjectivity is tasked with understanding how those sub-personal mechanisms generate personal-level content. Clever experiments probe the operating conditions under which the personal-level abstractions leak and the sparse implementation is revealed. But crucially, under “normal” operating conditions, the virtual subjective world is a robust abstract space of rich egocentric content, a functionally characterised one whose content we implicitly sample from whenever we introspect on our own subjective state – but which is not represented not by some kind of isomorphic data structure in our brain.³

3 De Facto Privacy and Heterophenomenology

The previous section sketched how the idea of virtualisation may provide a route to reconciling the apparent density or fullness of phenomenal content with the sparser story emerging from cognitive neuroscience. Here we suggest that virtualisation may help us understand the apparently private nature of subjective experience as well. This question has attracted less attention in the literature; many materialists, most notably Dennett [12], dismiss the idea as mistaken (Cartesian) folk theorising.

But this seems too quick. Construing our subjective lives as a kind of “inner observation” is a robust feature of our self-theorising, present even in small children [6]. It recurs in the various *problem intuitions* set out in Chalmers’ formulation of the “meta-problem” of consciousness [9]. These express familiar convictions: there are aspects of our subjective lives that are in-principle unknowable from the outside [18]; we can never know whether another creature – or a machine – is truly conscious (the so-called *Zombic Hunch* [11]); there are aspects of consciousness that are non-functional or epiphenomenal.

Given the prevalence of these so-called problem intuitions, a scientific account of subjectivity should arguably account for these as well. It may even be inaccurate to view them as problematic; perhaps we should instead see them as characteristic of the architecture of subjectivity. On Metzinger’s view, privacy is a consequence of what he calls the *transparency* of the self-model [22]. The self-model represents the agent as having a phenomenal life, but neither the existence of the self-model, nor its role in attributing phenomenal properties to the

³ This perspective is also consistent with a broadly *enactivist* understanding of subjective content, e.g. [2].

agent, is apparent to the agent itself. The agent thus interprets the properties assigned to it in the self-model *as* its intrinsic properties.

Virtualisation of subjective content offers a way to unpack Metzinger’s idea in terms of the distinction between the producer and consumer of content that we set out in Sect. 2.1. There we suggested thinking of the subjective agent as the consumer of (personal-level) phenomenal content, and sub-personal inferential mechanisms as delivering that content on an as-needed (virtual) basis. In such an architecture, an *abstraction boundary* separates, on the one hand, the personal-level agent (which need do no more than consider or introspect on its phenomenal world to initiate the generation of the required content), and on the other, the implementation mechanisms serving up the content in response to introspective queries. The abstraction boundary allows the personal-level agent, under normal operating conditions, to remain oblivious to all the sub-personal inferential activity that fixes and delivers content; from the vantage point of the agent, the content is seamlessly and continuously present. Indeed it is precisely the seamless and continuous presence of first-personal content that equips the agent *with* a vantage point in the first place.

So the transparency here arises because the subjective agent sits on one side of a functional abstraction boundary. On that side of the abstraction, information about *how the world seems* (or more carefully, *how the agent takes the world to seem*) from its perceptual vantage point is represented as various intrinsic, internal properties of the agent, similar to how the content of the matrix at coordinate i, j is “intrinsic” to the matrix from the vantage point of a user of the matrix abstraction, who remains oblivious to how that content is retrieved. As far as the agent is concerned, the properties ascribed to it in its self-simulation *are* its properties in the world, and so the agent takes itself to have these internal, intrinsic phenomenal properties. Determining “what it is like” [25] to be that agent is merely a matter of (effortless) private reflection, revealing rich and seamless content that the agent can decide whether (and to what extent) to share with the public world.

Thus virtualisation offers a concrete architectural perspective on Metzinger’s idea of transparency. Here it arises as a consequence of an abstraction boundary separating the production and consumption of the self-simulation – a boundary one might expect to see in systems with artificial subjectivity too. Moreover this is a putative architecture which is distinctively – although, we suggest, not problematically – Cartesian [26]. Because our (sub-personally self-assigned) phenomenal properties are, at the personal level, effortlessly and reliably present, we take ourselves to be *passive observers* of our own consciousness. In Hofstadter’s words, we “watch ourselves watching the world” [17].

3.1 The Fallacy of Non-Functionality

However private or non-functional we might naively take our phenomenal content to be, it is worth emphasising that such content is still a robust feature of our observable behaviour. To a certain extent our intuitions (paradoxically perhaps) pull in this direction as well: when we talk about our subjective lives,

we take ourselves to be doing so *because* we are conscious; we take ourselves to be *reporting on* the content of subjective experience. So we do seem to implicitly tie the observable behaviour – the reporting – to the presence of subjectivity. (Otherwise why even think of it as “reporting”?)

Indeed, as the philosopher Kim has pointed out [19], non-functional intuitions about consciousness are simply inconsistent with the idea that such things can be talked about in the first place. Any epiphenomenal view of consciousness falls foul of this fallacy. An epiphenomenal perspective supposes that there are at least some aspects of subjectivity (call them X) that are truly non-functional. But then, by assumption, no discourse which purports to be about X (call it “ X -discourse”) can ever arise as a consequence of the existence of X ; all such discourse is functionally independent of X and would carry on regardless. Any framing of the problem of consciousness in terms of epiphenomenality is therefore self-defeating: the only aspects of consciousness that X -discourse has any prospects of getting traction on are those aspects which are disjoint from the putative X .

Consider the vast number of colours we can discriminate [28] but are unable to describe effectively using language. Or the space of olfactory discriminations in humans, which seems to be even richer [7]. The significant shortfall in our reporting capabilities with respect to colour or olfactory phenomenology might lead us to think that there are subjective discriminations with no behavioural consequences at all.

But this is incoherent. While it is certainly possible, indeed common, for something we are *not* aware of to influence our behaviour, the opposite situation seems to be conceptually incoherent: the very presence of subjectively discerned details *constitutes* a capacity to behave about those discriminations – perhaps to prefer one stimulus over another in an experiment, or to produce one kind of verbal report over another (however coarsely related to the discrimination). While what we can distinguish subjectively significantly outruns what we can easily describe using language, this is at best a *de facto* limitation: all such discriminations afford some kind of overt action. If we can always extract information from the virtual subjective space by a suitably chosen “behavioural probe” [12], then all phenomenological data is heterophenomenological data.

4 Evaluating Machine Consciousness

So the claim is that we *behave as though* we were self-theorising agents, assigning rich subjective lives to ourselves which we take to be broadly private. We now close with some thoughts on how we might evaluate an artificial system to see whether it implements this sort of architecture. This is not as straightforward as opening up the machine and looking at the mechanisms inside, but not because subjectivity is fundamentally unobservable. Rather, the question of whether a system has this “design” is an empirical question about its *behaviour*, just as it is for us. With the virtual matrix, we cannot open up the computer and expect to find a 2D matrix inside, but must instead systematically probe the content of

the abstract representation by eliciting behaviour from the system; so it is with determining whether a system engages in this kind of self-theorisation. Here we briefly present some tentative empirical criteria that might support such an interpretation.

Ownership of First-Person Perceptual Vantage Point. Does the artificial system behave as though the world *seems a certain way to it* from its perceptual vantage point? In particular, is the agent able to distinguish how things *seem* to it (perceptually) from how it takes the world to *be* (given its current knowledge)? This seems to be characteristic of perceptual phenomenology; visual phenomenology, for example, allows us to distinguish between *how things look* from *how we take them to be*, allowing these two things to come apart, as exemplified by visual illusions. It seems plausible that an agent with visual experience, as opposed to mere visual perception, would be able to *appreciate*, as well as fall victim to, a visual illusion, and would be able to report on parallax, apparent shape and other perspectival information [26]. Consider how bats use Doppler shift to track relative velocity of prey. Our auditory phenomenology allows us to *appreciate* Doppler shift, such as when an ambulance passes by and the apparent pitch of the siren changes. But noting that bats exploit that information in predation is quite different from supposing a bat to have a personal-level appreciation that its auditory system is making that discrimination. These higher-order inferences are likely to be characteristic of systems with perceptual awareness [14].

Capacity for Deliberative, Reflective, Covert Behaviour. There is a large body of evidence that subjective awareness is important for sophisticated action [4]. Does the artificial system understand itself as situated in a world offering various affordances for action, many of which are covert? One of the strong reasons we can be certain that modern Large Language Models are not conscious is precisely because they fail to exhibit any behaviours consistent with a reading of them as engaged in this kind of self-simulation or self-theorisation. If prompted appropriately they can generate text strings superficially consistent with that possibility, but not engage in sustained reflective or deliberative behaviour.

Susceptibility to the Non-Functionality Fallacy. An intriguing possibility is that systems with artificial subjectivity may find the Hard Problem, if not convincing, then at least intuitively compelling. Does the agent make the same mistakes as we do about the private nature of subjectivity? Does it too appear to have instincts that pull in different directions? Does it seem plausible to the agent that certain behaviours in other organisms – for example cradling of injured limbs in cephalopods [1] – is evidence of phenomenal states? Is it skeptical of our ability to know for sure whether other agents are “really” conscious? These proclivities would all be consistent with the agent having a model of itself which assigns to the agent phenomenal states similar to ours.

None of the above is likely to constitute a conclusive test of subjectivity, but not because it is somehow fundamentally unknowable what goes on “inside the mind” of another agent. A conclusive test is unlikely just because precisely *how*

and *in what respects* an agent is conscious is empirical, multifaceted question. It comes down to whether we can make sense of the system's behaviour by attributing to it something resembling a virtual Cartesian theater. If this attribution – deploying as an empirical posit the hypothesis that “this agent has a subjective life” – yields a reliable way to explain and predict its behaviour, then that is the only basis for assuming that it does.

5 Conclusion

Building a conscious machine means making a machine that thinks it is conscious. We are such machines, albeit biological ones. For this to be a plausible picture, taking oneself to be conscious must not presuppose being conscious in the first place. But there is no reason for thinking that sub-personal mechanisms cannot account for personal-level content.

Taking oneself to be conscious is much richer than a mere self-directed propositional attitude: it must be closer to a continuous, egocentric *self-simulation* of the agent embedded in the world, ascribing to the agent rich phenomenal states, as envisaged by Metzinger and others. In this paper our starting point was the naturalistic idea that the explanatory task facing a science of subjectivity is to reconstruct the (ostensibly) interior subjective world of the agent using only the resources of “exteriority”, by operationalising subjective content as the capacity for certain kinds of behaviour, both covert and overt. Our main contribution was to show how the concept of *virtualisation* and the related distinction between abstract content and implementation mechanisms can shed light on the implementation of subjectivity in a way that preserves key realist intuitions in this naturalistic setting.

First, we argued that virtualisation allows room for the subjective *content* of the self-simulation to be dense but for the implementation of that content to be sparse; and there are good reasons, both theoretical and practical, for expecting the implementation to be sparse in both natural and artificial systems. Much of the cognitive neuroscience data on sparse representations should be interpreted as evidence for this style of implementation, not as evidence that the density of personal-level content is an illusion. Second, we revisited Metzinger's idea of transparency and the relationship to the (apparent) privacy of subjective content. Here the idea of an abstraction boundary isolating the personal-level agent from the implementation mechanisms delivering phenomenal content made it clearer how, in a functional sense, we do plausibly “observe ourselves” being conscious. Functionally, a conscious agent inhabits a “virtual” Cartesian Theatre, where phenomenal content is fixed on an as-needed basis as the agent queries its own experiential state. There is a *de facto* notion of privacy because such content is seamlessly and immediately present to the agent, but no genuinely private content.

References

1. Alupay, J.S., Hadjisolomou, S.P., Crook, R.J.: Arm injury produces long-term behavioral and neural hypersensitivity in octopus. *Neurosci. Lett.* **558**, 137–142 (2014). <https://doi.org/10.1016/j.neulet.2013.11.002>
2. Anderson, M.L., Rosenberg, G.: Content and action: the guidance theory of representation. *J. Mind Behav.* **29**(1/2), 55–86 (2008)
3. Baars, B.J.: Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. In: Laureys, S. (ed.) *The Boundaries of Consciousness: Neurobiology and Neuropathology*, Progress in Brain Research, vol. 150, pp. 45–53. Elsevier (2005). [https://doi.org/10.1016/S0079-6123\(05\)50004-9](https://doi.org/10.1016/S0079-6123(05)50004-9)
4. Baumeister, R.F., Lau, S., Maranges, H.M., Clark, C.J.: On the necessity of consciousness for sophisticated human action. *Front. Psychol.* **9** - **2018** (2018). <https://doi.org/10.3389/fpsyg.2018.01925>
5. Blanke, O., Metzinger, T.: Full-body illusions and minimal phenomenal selfhood. *Trends Cogn. Sci.* **13**(1), 7–13 (2009). <https://doi.org/10.1016/j.tics.2008.10.003>
6. Bloom, P.: *Descartes' Baby: How The Science Of Child Development Explains What Makes Us Human*. Basic Books (2009)
7. Bushdid, C., Magnasco, M.O., Vossell, L.B., Keller, A.: Humans can discriminate more than 1 trillion olfactory stimuli. *Science* **343**(6177), 1370–1372 (2014). <https://doi.org/10.1126/science.1249168>
8. Chalmers, D.: *The Conscious Mind*. Oxford University Press, Oxford (1996)
9. Chalmers, D.J.: The meta-problem of consciousness. *J. Conscious. Stud.* **25**(9–10), 6–61 (2018)
10. Dehaene, S., Kerszberg, M., Changeux, J.P.: A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci.* **95**(24), 14529–14534 (1998). <https://doi.org/10.1073/pnas.95.24.14529>
11. Dennett, D.: The zombic hunch: extinction of an intuition? *R. Inst. Philos. Suppl.* **48**, 27–43 (2001). <https://doi.org/10.1017/S1358246100010687>
12. Dennett, D.C.: *Consciousness Explained*. Little, Brown (1991)
13. Dennett, D.C., Kinsbourne, M.: Time and the observer: the where and when of consciousness in the brain. *Behav. Brain Sci.* **15** (1992)
14. Fleming, S.M.: Awareness as inference in a higher-order state space. *Neurosci. Conscious.* **2020**(1), niz020 (2020). <https://doi.org/10.1093/nc/niz020>
15. Frankish, K.: Illusionism as a theory of consciousness. *J. Conscious. Stud.* **23**(11–12), 11–39 (2016)
16. Graziano, M.: The attention schema theory: a foundation for engineering artificial consciousness. *Front. Robot. AI* (2017). <https://doi.org/10.3389/frobt.2017.00060>
17. Hoenderdos, P.: *Victim of the brain* (documentary). <https://www.imdb.com/title/tt0096382/> (1988)
18. Jackson, F.: Epiphenomenal qualia. *The Philosophical Quarterly* (1950) **32**(127), 127–136 (1982)
19. Kim, J.: *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press (1998)
20. Lampson, B.: Hints for computer system design. In: *9th ACM Symposium on Operating Systems Principles*, pp. 33–48. ACM, ACM (1983)
21. Metzinger, T.: *The Ego Tunnel: The Science of the Mind and the Myth of the Self*. Basic Books (2009)
22. Metzinger, T.: *Being No One: The Self-Model Theory of Subjectivity*. MIT Press (2003)

23. Metzinger, T.: Précis: Being No-One. *PSYCHE: Interdis. J. Res. Consci.* **11**, 1–30 (2005)
24. Metzinger, T.: Empirical perspectives from the self-model theory of subjectivity: a brief summary with examples. In: Banerjee, R., Chakrabarti, B.K. (eds.) *Models of Brain and Mind*, Progress in Brain Research, vol. 168, pp. 215–278. Elsevier (2007). [https://doi.org/10.1016/S0079-6123\(07\)68018-2](https://doi.org/10.1016/S0079-6123(07)68018-2)
25. Nagel, T.: What is it like to be a bat? *Philos. Rev.* **83**(October), 435–50 (1974)
26. Perera, R.: Cartesian creatures: watching ourselves watching the world. *J. Conscious. Stud.* **26**(3–4), 131–154 (2019)
27. Pessoa, L., Thompson, E., Noë, A.: Finding out about filling in: a guide to perceptual completion for visual science and the philosophy of perception. *Behav. Brain Sci.* **21** (1998)
28. Pointer, M.R., Attridge, G.G.: The number of discernible colours. *Color Res. Appl.* **23**(1), 52–54 (1998)
29. Ryle, G.: *The Concept of Mind*. Hutchinson & Co (1949)
30. Ryle, G.: The thinking of thoughts: What is ‘Le Penseur’ doing? *University Lectures* **18** (1968)
31. Schiller, P.H.: The ON and OFF channels of the visual system. *Trends Neurosci.* **15**(3), 86–92 (1992). [https://doi.org/10.1016/0166-2236\(92\)90017-3](https://doi.org/10.1016/0166-2236(92)90017-3)