Research Note (version 2 August 2021)
An updated version of this paper can be found as Chapter 16 ('Philosophical problems of consciousness') in my book 'The estimator theory of life and mind: how agency and consciousness can emerge', see https://philpapers.org/rec/VANTET-8

Solutions to some philosophical problems of consciousness

J. H. van Hateren
University of Groningen, The Netherlands, j.h.van.hateren@rug.nl

Abstract A recently developed computational and neurobiological theory of phenomenal consciousness is applied to a series of persistent philosophical problems of consciousness (in recent formulations by Tye, Searle, and Chalmers). Each problem has a clear solution according to this theory, as is briefly explained here.

Philosophical analysis of consciousness has produced a rich literature on actual and potential problems of consciousness. Several major problems—partly derived from formulations by Tye (2017), Searle (2017) and Chalmers (2017)—are analysed below from the perspective of a recently proposed theory of consciousness (van Hateren 2019) and intentionality (in the sense of 'aboutness', van Hateren 2021a). Section 1 summarizes the theory, but the reader is advised to consult the abovementioned papers as well as the one on strong emergence (van Hateren 2021b). Although the theory is conjectural and requires empirical investigations, I will avoid the many mays and mights that could litter the text below. Instead, it is written as if the theory concerns established facts. But the reader should keep in mind that such is merely a stylistic choice.

A note on terminology and notation may be helpful. The term 'intentional component' is used below for what is called X_i in van Hateren (2019) and X-component in van Hateren (2021a). Similarly, 'fitness component' corresponds to F_i (or F-component), and inverted intentional component (which is experienced) corresponds to $\bar{X}_i$. The estimate of an individual's fitness is denoted by x both here and in van Hateren (2019, 2021a,b), but by f_{est} and $\hat{f}$ in van Hateren (2015a,b), where the process X that produces the value x is denoted by the form of f_{est} (analogously to a mathematical function that has both a value and a form). A similar notation concerns fitness f_{true} and f (van Hateren 2015a,b), which is denoted by a process F that produces a value f in van Hateren (2019, 2021a,b).

1 Summary of the theory of consciousness

A key feature of any biological organism is its evolutionary fitness f, which is, in the simplest form, its propensity to survive and reproduce—the term 'propensity' indicates that fitness is used here as a forward-looking, predictive factor. However, fitness is often not so simple, because it can include the effects of helping related individuals (which is known as inclusive fitness), as well as social and cultural effects. Under quite general conditions (such as on heredity and variability of traits), fitness differences between individuals lead to evolution by natural selection.

Fitness as defined here acts continuously during the lifetime of any organism. Therefore, each organism typically strives to keep its fitness high during its lifetime, through various mechanisms. Usually, such mechanisms are primarily deterministic, but it is in fact possible to

enhance fitness through a remarkable non-deterministic mechanism (van Hateren 2015a). An organism then produces an implicit internal estimate, x, of its own evolutionary fitness, and utilizes that estimate when randomly varying its internal structures (with x and its effects present in a distributed way). The variation is done in the following way: when fitness is estimated to be low, structures are changed with much variability ('desperate times call for desperate measures', if desperate includes undirected), whereas a high fitness estimate produces little variability ('never change a winning team', or at least not much). This mechanism is conjectural but can be shown to be evolvable (see computations in van Hateren 2015a). The key point here is that evolution can produce estimation as a causal factor: the better the internal estimate of fitness is, the higher the subsequent fitness (denoted by 'fitness-to-be') will become—slowly and gradually because the mechanism is stochastical.

Estimation does not exist as a causal factor in abiotic nature; thus, it is a purely biological novelty. It can be regarded as a minimal form of intentionality (van Hateren 2021a). Importantly, this particular form of estimation can be shown to be a strongly emergent cause (van Hateren 2021b). This means that its causal efficacy cannot be explained by any set of micro-causes (essentially because it depends in a cyclical way on the structural effects of noise). As a result, estimation exists in a literal sense, as a distinct and autonomous entity. However, this entity is not well localized, because what is estimated, fitness, is produced by a complex process F with components that are scattered widely throughout the world.

An entity that is well localized to the brain emerges—again in a strongly emergent way—when components of the fitness estimate (called 'intentional components' below, see van Hateren 2021a) are being prepared to be communicated to a related organism. If the setting is assumed to be cooperative, the inclusive fitness-to-be of the sender will increase, on average. Preparing to communicate an intentional component by a sender requires—in its most basic form—an inversion, such that it leads to a similar intentional component in the receiver (this depends on the fact that an operation followed by its inverse produces an identity operation). Inversion can be performed by the thalamocortical feedback loop in the mammalian brain, if it is used in a switched, dual-stage way (van Hateren 2019). The first stage produces intentional components, whereas the second stage inverts them through a specific feedback mechanism. Stages are switching continually, at a rate of roughly 10 Hz in the primate brain. Inverted intentional components are either communicated to a partner or are used internally as further input to the thalamocortical loop.

Inverted intentional components are causal factors that can be shown to be both strongly emergent and spatially localized (van Hateren 2019, pp. 367–369). They produce an entity that is autonomous, distinct, spatially localized to the brain, transient, and strongly emergent. Thus, they appear to mimic a localized material cause (i.e., as if they were a transient material object within the brain) and it is plausible that their presence is sensed, as the feeling of consciousness. Their content equals that of the corresponding intentional components. The total content of consciousness depends on which inverted intentional components are active at any point in time. The unity of consciousness is produced by the fact that all intentional components get their causal efficacy from the causal efficacy of the overall estimate x of fitness f, which are both scalars (and thus unitary).

The above summary focusses on the most basic form of consciousness, which occurs when intentional components are being prepared to be communicated. If this happens internally, it can set up an internal conscious cycle. Consciousness can be produced by perception in an indirect way, such as when a communicated intentional component or a visual scene induces an internal conscious cycle. Complex forms of consciousness can arise in a way that is similar to how complex forms of intentionality can arise (see van Hateren 2021a). The relationship

between intentionality and consciousness is indeed a close one: consciousness arises when intentionality is being prepared to be communicated.

2 Problems and solutions

2.1 Ownership (Tye 2017, p.18)

Problem. Specific subjective experiences are necessarily owned by a specific individual. This makes them different from ordinary physical things. Such things are sometimes owned, but they can still exist if they are not owned. In contrast, subjective experiences cannot exist in an unowned state. Thus, there is a problem if one assumes that phenomenal consciousness is wholly physical.

Solution. Although consciousness is produced by a physical system, it depends on strongly emergent causal factors, specifically estimation and its inversion. Inversion of estimation produces a concrete entity (in the sense of being distinct and spatially localized) that is sensed, but that is *not* an ordinary physical thing (in the sense of consisting of matter-energy). It is attached inseparably to the neurobiological system that produces and owns it, and it cannot exist in an unowned state.

2.2 Perspectival subjectivity (Tye 2017, p.19)

Problem. Phenomenal conscious states are perspectival, but physical states are not. Whereas the latter can be fully understood from a complete description of state and dynamics, the former can only be fully comprehended by having the proper experiential perspective (such as when having a pain, feeling a depression, and having the visual experience of red).

Solution. A conscious state (or, more appropriately, a specific conscious process) consists of components that are the inverse of specific intentional components. Thus, the content of conscious experience depends on the content of intentionality. The latter is a form of estimation produced by the brain. It concerns an estimate of components of the individual's own evolutionary fitness, produced by the individual itself. Hence, it has a subjective perspective. It is not an ordinary physical state, because it is a strongly emergent entity.

2.3 Mechanism (Tye 2017, p.20)

Problem. What is the mechanism that produces the what-it's-like feeling? In the natural world, it seems that higher-level states or processes or properties are always grounded in—and are explained by—what is going on at lower neurophysiological or chemical or microphysical levels.

Solution. Ontological reduction may be generally applicable in the abiotic natural world, but not here. The key point here is evolutionary fitness, which confers causal efficacy (through affecting fitness-to-be) to an internal fitness estimate made within the individual. This mechanism depends in an indispensable way on the micro-effects produced by randomness. Estimation is a novel and strongly emergent causal factor that has thus been added to nature (by having evolved through natural selection). The same applies when estimation is inverted for the purpose of communication. The existence of strongly emergent causes implies that the physical

world is not completely causally closed (van Hateren 2021b). Applying philosophical concepts like 'grounding' and 'supervenience' to the system that produces consciousness is problematic, because ontological randomness across a stretch of time does not correspond to physical 'facts' (van Hateren 2021b). Even 'what is going on' would be undefined in all its details (unless one tacitly and falsely presupposes determinism).

2.4 Duplicates 1 *(Tye 2017, p. 21)*

Problem. A philosophical zombie is taken to be a perfect material duplicate of a conscious being, except that it completely lacks phenomenal consciousness. Otherwise, it has identical behaviour and identical mental processes. Usually, it is not claimed that such zombies are physically possible (given the features of the world that is), but rather that they are imaginable or logically possible or metaphysically possible. If they are, then consciousness seems separable from its material substrate.

Solution. Consciousness is not separable from its material substrate, and philosophical zombies are therefore not possible—neither with 'possible' in the sense of feasible, nor in the sense of conceivable, nor in the sense of non-self-contradictory. Once there is a valid and accepted explanation of how consciousness arises, one is not free any more to use one's imagination or logic or metaphysical assumptions in a way that conflicts with that explanation. That would amount to basing an argument on false or implausible premisses. Thus, arguments based on philosophical zombies are unlikely to be sound.

2.5 Duplicates 2 *(Tye 2017, p.22)*

Problem. One might simulate the brain in arbitrarily fine detail in another system, such as might be realized by one billion carefully instructed people. Intuitively, one would think that such a system (as a whole) would not be conscious, even if it would perform a perfect simulation.

Solution. Perfect simulation is not possible in this way. The main problem with this kind of simulation is that there can be no internal estimate of fitness (which is required for modulating random structural change in the system) because there is no fitness to estimate. One billion people do not survive and reproduce as a unitary entity (such as by multiplying at once to two or three billion, or dying at once to zero). Moreover, there is no competition and cooperation with other such entities in a shared environment, nor a well-defined, unitary heredity of structure; therefore, there is no evolution by natural selection. Estimation and intentionality depend on sustained evolution by natural selection. Without real estimation and intentionality, there can be no strongly emergent and distinct entity that is felt as consciousness.

2.6 The inverted spectrum *(Tye 2017, p.23)*

Problem. Suppose that Tom has a very peculiar visual system (perhaps produced by a neurosurgical rewiring at birth), such that he experiences red where others experience green, and vice versa. But nobody is aware of this difference, because otherwise Tom functions as anybody else. Thus, there is a phenomenal difference without a functional difference. More generally, one may suppose that such phenomenal inversion can occur even in microphysical duplicates.

Solution. Subjective experience is the sensed entity that is produced by inverting intentional components. Hence, the quality of the experience depends on the content of the corresponding intentional components. The content of an intentional component depends on the fitness component that it estimates, including the role that this component has in the process that produces fitness. The colour red has approximately the same fitness associations in a group of culturally and functionally similar people (think of typical red things: strawberries, sunsets, fires, roses, traffic lights, socialism, blood, and so on). Therefore, their intentional components concerning red are roughly similar, and thus is their phenomenal experience of red. This is no different for Tom and his peers. The assumption that Tom is possible (in any sense of the term) is false. Any rewiring at birth would still produce the same phenomenal experience if Tom indeed functions as anybody else.

2.7 Transparency *(Tye 2017, p.24)*

Problem. When attending to a visual experience, one becomes aware of what is seen (such as a particular object and its qualities), but not of the experience as such. Thus, phenomenal consciousness seems to be transparent. Why, then, is it felt?

Solution. It is felt because it equals a distinct, strongly emergent, transient, and spatially localized entity (which is identical to the strongly emergent causes produced by inverting intentional components). The content of this entity is the content of the corresponding intentional components (pointing to a particular object and its qualities). Thus, the entity has no additional content. Having no additional content may be interpreted, incorrectly, as transparency. The interpretation is incorrect, because entity and content are not separable.

2.8 Unity *(Tye 2017, p.25)*

Problem. There is a unity to conscious awareness. The different items that make up a specific conscious experience (e.g., the perceived objects, actions, and sensory impressions in a particular setting) are not experienced as fully separate. Rather, they are perceived as integrated in the whole. Similarly, conscious experiences stay integrated across time. How can that be?

Solution. Consciousness at any time consists of a large set of inverted intentional components. Their content corresponds to the content of the corresponding intentional components. Intentional components estimate fitness components (aspects of an individual's fitness) in such a way that together they produce a unitary (scalar) estimate of the individual's fitness, x. This estimate has strongly emergent causal efficacy, which is, ultimately, the reason why consciousness is felt. The intentional components (as well as their inverted versions) are automatically integrated by x. This is not only true at any point in time, but also across time, because X, the process that produces x, is maintained across time (even as it changes gradually).

2.9 Divided consciousness *(Tye 2017, p.27)*

Problem. In split-brain patients the corpus callosum is cut (for medical reasons), which drastically reduces the communication between left and right half of the cortex. When conflicting information is presented to the left and right half of a patient's brain, perception seems to occur locally, without being communicated to the other half. Thus, perception is divided. Does such a patient, then, have a split consciousness too?

Solution. Consciousness is conjectured to be produced by the second stage of a dual use of the corticothalamic feedback loop (van Hateren 2019). This second stage inverts intentional components that are presumably produced by a wider loop involving thalamus, cortex and basal ganglia, with important inputs from the upper brain stem. Together these establish x, the (distributed) estimate of an individual's evolutionary fitness. Specific parts of the left or right cortex are then participating in specific intentional components, corresponding, for example, to specific visual perceptions. However, the unity of consciousness itself does not fully depend on the unity of left and right cortex. It also depends on the left-right unity of thalamus, basal ganglia and upper brainstem (as these produce x too). The latter unity remains intact in split-brain patients. Hence, there is no reason to assume that these patients have a fully split consciousness. Moreover, they are still one individual with one fitness, thus they are likely to learn compensating strategies that repair the unity of their x, even if it were compromised initially. This may explain why split-brain patients still feel as one.

2.10 Animal consciousness (Tye 2017, p.28)

Problem. How can we decide which other creatures have consciousness?

Solution. For consciousness, creatures need to have evolutionary fitness and need to make an internal estimate of that fitness (which then stochastically drives structural changes in the creature's brain). Moreover, they need to invert components of this estimate, in preparation for internal or external communication. The capacity to communicate intentionality to conspecifics in a cooperative setting (thus typically increasing inclusive fitness) must be present at least, as a basis for more elaborate external or internal communication. Then inverting estimated fitness components produces strongly emergent causes, which constitute the distinct, strongly emergent entity that is felt as consciousness. In summary: in order to have consciousness, creatures must have evolutionary fitness, an internal fitness estimate driving a specific stochastic mechanism, inversions of this estimate's components, and cooperative communication of intentionality with at least one conspecific. These conditions are sufficient, and they are in principle amenable to empirical assessment through neurophysiological and behavioural research.

As a poor man's test of consciousness, one may try to engage a creature in a dialogue of (nonverbal) intentionality, thus establishing some sense of mutual rapport. When establishing a mutual empathic bond is easy (as with mammals and birds), this indicates the presence of consciousness, and when this seems impossible (as with worms and even with social insects), this indicates its absence.

Note that the above considerations assume a specific mechanism for producing consciousness. Although it is highly specific and may well be the only one capable of producing consciousness, it cannot be ruled out—at this point in time—that alternative mechanisms exist.

2.11 Causal efficacy (Searle 2017, p.330)

Problem. One can initiate behaviour by a conscious decision. How is that possible if the brain is fully functioning through neural mechanisms?

Solution. It would indeed not be possible if neural mechanisms were deterministic or at least were ontologically reducible. But the neural mechanism that produces consciousness is neither.

Consciousness consists of sets of inverted intentional components that can be used as internal input to produce intentional components, which are subsequently inverted and then used as internal input once more, and so on (van Hateren 2019). The totality of intentional components changes through time in this way, which is equivalent to a change of the structure of the X process that produces x. The latter drives random structural changes in the brain, including ones that affect behavioural dispositions. Behavioural dispositions that produce large x appear to be sticky (because large x produces a low rate of structural change), whereas behavioural dispositions that produce small x appear to be repellent (because small x produces a high rate of structural change). Which particular behavioural dispositions produce small or large x is determined not only by the input to the X process but also by its structure, and is, thus, partially controlled by how consciousness proceeds. Hence, consciousness affects which behavioural dispositions are present. Therefore, it affects the resulting behaviour—albeit by a slow, stochastic process. Instant behavioural decisions need to be prepared in advance, as stored dispositions that can be utilized nonconsciously or at least preconsciously (see also van Hateren 2015b).

2.12 Dancing qualia (Chalmers 2017, p.369)

Problem. Two functionally isomorphic systems must have the same sort of experiences. For example, a conscious biological organism may be gradually replaced, neuron by neuron and cell by cell, by silicon equivalents (this is utterly unrealistic[1], but let us suppose that it could be done). If one claims that the final, silicon version has different consciousness, or no consciousness at all, then there might be, at some point along the transition, a significant shift in experience. Moving back and forth across this point would produce dancing qualia (qualities of experience). This seems counterintuitive, thus functional isomorphism must imply equal subjective experience.

Solution. Replacing biology by silicon may not leave fitness intact, that is, the final silicon version may have lost the capacity to reproduce and the propensity to die. If that is so, the silicon version cannot make an internal fitness estimate (other than a fake one that would quickly fall short). Even if the silicon version had fitness (the propensity to survive and reproduce) it would not have inclusive fitness if it were the only one of its kind. Then inverting intentional components would not be sustainable (for lack of inclusive fitness), and neither would be consciousness. Assuming that fitness is indeed lost, the thought experiment would not show a sharp transition between the presence and absence of consciousness. Rather, the silicon version would gradually loose more and more of its consciousness when it senses—implicitly or explicitly—that it is getting more and more alienated from its former conspecifics. Being indefinitely alone in the world, without any prospect of a meaningful future, is not consistent with sustaining consciousness.

2.13 Machine consciousness

Problem. Can machines become conscious?

[1] Neurons work and communicate, as all biological cells, at a molecular level; it is difficult to see how such specific processes could be replaced by processes with a different material basis without producing considerable consequences for fitness. What about the mass, energy requirements and volume of the replacement? Heat dissipation? Functional noise? Structural changeability? Reproduction? Repairability? Molecular defences against disease? And so on and so forth.

Solution. Short answer: no, unless machines become alive first; but it is doubtful (or at least a definitional issue) if one could still call such a living system—an organism—a machine. Long answer: consciousness arises in a system when intentionality is transformed such that it can be communicated (externally or internally) and thus can increase inclusive fitness. Intentionality is a strongly emergent phenomenon that depends on a fitness estimate that modulates random structural change of the system. This mechanism is only sustainable when the fitness estimate accurately estimates a real fitness, with real reproduction (because the exponential growth of reproduction is needed in order to compensate for the inefficiency of random structural change; this ultimately depends on evolution by natural selection). Fitness, random structural change, and evolution by natural selection are defining features of life. Therefore, a system needs to be alive in order to have intentionality and subsequently consciousness. Whether a living system could be called a machine is debatable. In any case, building such a system would be risky, because it would try to replicate without bounds, and would thus compete with humans and other biological life forms.

2.14 How could having consciousness produce evolutionary benefits?

Problem. Apparently, having consciousness is an evolved property in some species. If so, which evolutionary advantages would it confer on these organisms?

Solution. Consciousness is not a trait that can be separated from the mechanism that produces it as a strongly emergent entity. Thus, the evolutionary advantages of consciousness are equal to the evolutionary advantages of this mechanism. The mechanism is the transformation of intentional components into a form that can be communicated, at the very least to conspecifics that are inclined to cooperate. Communication of intentionality will then increase inclusive fitness, on average. Therefore, this transformation is evolvable, and the strong emergence of consciousness is the automatic consequence. Note that this does not make consciousness an epiphenomenon, because it is identical to the occurrence of the transformation. The accuracy of the transformation is a strongly emergent entity with causal power, and, thus, by no means an epiphenomenon. In summary, the evolutionary benefits of consciousness are identical to the evolutionary benefits of transforming intentional components for communication. The latter enhances inclusive fitness, on average.

2.15 Wouldn't a fully non-communicative species still benefit from experiencing pain?

Problem. If consciousness is, in its most basic form, communicative rather than perceptive, wouldn't this imply that a species that has no use for communicating intentionality has no consciousness? But wouldn't experiencing, such as experiencing pain, provide evolutionary benefits anyway?

Solution. If a species lacks the capacity to communicate intentionality, it has indeed no subjective experience. Stimuli or internal states that would indicate harm can then still lead to behaviour that alleviates the problem, but there would be no associated subjective experience. This is so, because such a species lacks a neuronal system that transforms intentional components for communication to others or for further internal processing. Experiencing is inseparably coupled to this transformation (as a strong emergent), and talk about the benefits of the experience as such makes no sense. Any benefits must arise from the prospective

communication to others, because benefits depend on the parts of inclusive fitness that go beyond direct (individual) fitness.

2.16 Consciousness and quantum physics both seem weird. Is there a link?

Problem. Is there a link between consciousness and quantum weirdness, such as entanglement and wave function collapse upon observation?

Solution. Indirectly. The theory of consciousness and intentionality depends on randomness that is ontological (thus 'out there' and not just a consequence of insufficient knowledge). The source of such randomness is thermal in practice, specifically in the form of random fluctuations of the fairly small number of molecules that are typically involved in (neuro)physiological processes. Such molecular randomness may ultimately depend on quantum randomness, because nonlinear dynamical systems can amplify submicroscopic fluctuations to microscopic and macroscopic ones. Quantum randomness appears to be fundamental. But apart from ontic randomness, there does not seem to be a link between other forms of quantum weirdness and consciousness.

A potential issue here is that the correct interpretation (or foundation) of quantum physics is not yet clear, with some interpretations seemingly suggesting determinism. If the theory of consciousness discussed here acquires empirical support, then full determinism becomes less tenable. If, on the other hand, a fully deterministic physics acquires empirical support, then this particular theory of consciousness becomes less tenable.

2.17 Could mind and consciousness be uploaded to a computer?

Problem. If one assumes that mind and consciousness are produced by some specific kind of information processing in the brain, shouldn't it be possible to upload the relevant information to a computer, and then simulate or emulate consciousness?

Solution. No, this is not possible. Consciousness is not produced by a specific kind of information processing. Rather, it requires a physical body that participates in sustained evolution by natural selection and that incorporates a causally effective internal estimator of its evolutionary fitness (see also the discussion of the 'brain in a vat' thought experiment in van Hateren 2021a). To the extent that neural processing can be described as information processing, this always concerns meaningful information. Such information is necessarily about something, and thus depends on intentionality. The assumption that information processing produces intentionality and consciousness, and can be used for explaining them, is viciously circular.

3 Conclusion

All problems discussed above have a clear solution if the proposed theory of consciousness turns out to be correct.

References

Chalmers D (2017) Naturalistic dualism. In: Schneider S, Velmans M (Eds) The Blackwell Companion to Consciousness (Second Edition), pp. 363–373. Wiley-Blackwell, Chichester

Searle J (2017) Biological naturalism. In: Schneider S, Velmans M (Eds) The Blackwell Companion to Consciousness (Second Edition), pp. 327–336. Wiley-Blackwell, Chichester

Tye M (2017) Philosophical problems of consciousness. In: Schneider S, Velmans M (Eds) The Blackwell Companion to Consciousness (Second Edition), pp. 17–31. Wiley-Blackwell, Chichester

van Hateren JH (2015a) Active causation and the origin of meaning. Biol Cybern 109:33–46 https://doi.org/10.1007/s00422-014-0622-6 (no open access) or arXiv:1310.2063

van Hateren JH (2015b) The origin of agency, consciousness, and free will. Phenom Cogn Sci 14:979–1000 https://doi.org/10.1007/s11097-014-9396-5 (no open access) or preprint

van Hateren JH (2019) A theory of consciousness: Computation, algorithm, and neurobiological realization. Biol Cybern 113:357–372 https://doi.org/10.1007/s00422-019-00803-y (open access)

van Hateren JH (2021a) Constructing a naturalistic theory of intentionality. Philosophia 49:473-493. https://doi.org/10.1007/s11406-020-00255-w (open access)

van Hateren JH (2021b) A mechanism that realizes strong emergence. Synthese, 2021, https://doi.org/10.1007/s11229-021-03340-z (open access)