On-off synecdoche: A just good enough model of subjective experience

Mario Martinez-Saito *

Institute of Cognitive Neuroscience, HSE University, Russian Federation

Abstract

We provide an account of the apparent intermittency of subjective experiences, between conscious and unconscious phases, grounded on three parsimonious notions: (1) the brain is an inference engine and stochastic simulator endowed with a good enough generative model of the world inherited via evolution and forged by experience, (2) the brain's internal model of the self (itself, its body and actions) is an intermittent and simplified representation invoked only when needed to expedite inference, which specifies and enables first-person subjective experiences through the identification of a model of the self (d-self or synecdoche for the p-self) with the physical self (p-self), and (3) realistic monism as a merger of physicalism and panpsychism. This scheme shifts the focus from problematic standalone subjective experiences to the identification of subjective experiences with the system's model of itself and its contingent attributes, is consistent with the empirical and phenomenological evidence and provides testable predictions: (1) only macroscopic scale information that is expedient for survival (d-self shell) can become subjective experience, and (2) the hub of subjective experience is mostly distributed along the posterior medial cortex.

Keywords: approximate inference; consciousness; self-model; realistic monism

Contents

*mmartinezsaito@gmail.com

1. Introduction: the illusion of reality

Theorizing about consciousness can easily become a wild goose chase. The unique standing of consciousness science straddling physics and phenomenology makes it a sitting duck for diverging opinions, as attested by its sprawling literature. This article lays out a theory about consciousness that attempts to eschew the many pitfalls lying on the path towards understanding consciousness and also to harmonize several of the most popular theories of consciousness into a single coherent scheme.

There are at least two aspects to the puzzle of consciousness: the nature of subjective experiences and qualia —i.e. explaining *all* their properties, including phenomenal attributes [203]— and the mapping between physical states and conscious states. The former is Chalmer's hard problem of consciousness [41]: how can third-person quantitative science explain the first-person aspects (phenomenology) of mental states? We do not provide a satisfactory answer to the hard problem. Instead, here we focus on the "easy" problem, particularly on the question: "What distinguishes conscious from unconscious brain states?"

Perhaps the major hurdle in unraveling the riddle of consciousness is the illusion of reality[1]. We tend to take for granted everything we perceive, especially what we see with our own eyes. However, going at least as far back as the 4th century BCE, some people, such as Zhuangzi in China [306] and the Pyrrhonists and the Academic Skeptics (e.g. Carneades) of Classical Greece [169], have suspected or asserted that there is no way to make sure that perception bears conformity to reality. Later, Hume's skepticism about the possibility to infer certainties about the world [144] would inspire Kant to expound in his epistemological doctrine of transcendental idealism [154] that we cannot directly cognize what lies outside our heads: whatever it is we are seeing, it is not the same thing that exists "out there" or *noumenon* but a synthesis of our built-in dispositions (e.g. internal generative model) with sensory input that brings about our inner experience —thus anticipating the importance of unconscious inference [129, 121, 101]. Yet the illusion that we perceive *directly* is so stubborn and persevering that, even today with the profusion of evidence indicating that perception is mostly a matter of guessing [202, 60, 95, 163, 285], we cannot stop ourselves from believing that we see as it is. This illusion of access to reality hinders our ability to relate phenomenal to physical properties because it bypasses the substrate of our experience, which is some subset of the brain.

The illusion of reality is particularly conspicuous in the contrast between dreaming and wakefulness, as recounted by Zhuangzi: "While he is dreaming he does not know it is a dream, and in his dream he may even try to interpret a dream" [306]. This leads to the conclusion (dream argument) that we lack even an elementary ability to judge

[1] Also called doxastic veridicality [249].

the conformity between internally generated representations and the actual state of the environment: by default, we believe whatever the brain arrives at. The illusion of reality pervades all levels of cognition, from sensation to the most abstract concepts.

Near the top of the cognitive hierarchy of ontologies, there is a singularly important concept: the self. The concept of self is essential to most living organisms because it follows from the need to discriminate between the organism and the environment. We will argue that the realization that we consistenly conflate the internal model of the self —realized as a neural and conscious representation— with the actual physical self —the actual physical state corresponding to the internal model of the self— may be the key to elucidate most of the "easy" problem of consciousness.

2. Everting consciousness theories

Although the concept of self is intuitively clear for most earthly purposes, it becomes ambiguous as soon as one starts prying its semantic boundaries. There is no formal definition of self, but roughly it refers to the body we own, the actions we choose, and our current, past, and future (or altogether *diachronic*, meaning "throughout time", see [274, 29] for a similar usage) attributes and states, as opposed to environmental objects or phenomena that are *not* our body, are *not* caused by us, and are *not* our attributes or states. A self and its environment are mutual complements with respect to the world.

Here, we will be concerned with two sorts of self: (1) the dummy self, deictic self, homunculus, "synecdoche" for the p-self, diachronic self or d-self: a living being's internal model of itself, typically implemented as a neural representation in the brain; and (2) the physical self or p-self: its (supposed) actual physical counterpart.

The common or folk psychology sense of subjective self[2]*, which endows us with selfhood through the belief of being an individual with its attending knowledge and feelings, is a conscious expression of the d-self.* Appreciating the significance of this observation requires spelling out what we mean by brain internal model, self (d-self and p-self), and consciousness, which is the purpose of Sections 3, 4, and 5.

One of the consequences of the illusion of reality is that we believe our self (d-self) state to be whatever the brain figured out its internal model of the self state is. Crucially, the d-self is just a *model* of the physical self or p-self. This implies that the outset in any inquiry is already an approximation to physical reality: it may be good enough for most purposes, but it is inaccurate, and sometimes plain wrong.

There is an essential difference between the brain internal representation of environmental objects such as say a brown onion bulb standing on a table and the representation of oneself (d-self) as an agent. Under reasonable circumstances, both entities are expected to continue existing in time for some extended interval. This expectation is easy to satisfy for the onion by simply assuming that it was still on the table even when I was not looking at it. But it is difficult to justify the time-continuity of the d-self when there are gaps in the memory where I do not know what I was doing (e.g. clinical unconsciousness, sleeping, mind-wandering, Section 4.3.2); this is because our conscious actions and subjective experiences, apart from informing the state of the world, are also integral components to the definition of the self when considered as choices and perceptions belonging to the d-self. This agrees with our belief that the p-self is in essence an entity persistent in time, even if we cannot account for this continuity on the basis of

[2]This roughly corresponds to the self-concept as defined in psychology.

our intermittent disjoint memory chunks of d-selves. In other words, the discontinuity of the subjective self or d-self is at odds with the presumed continuity of the physical self or p-self.

This leads us to the unorthodox approach of *everting* —meaning turning inside out— the puzzle of consciousness. This refers to turning on its head the common notion that I am an observer with conscious and unconscious or on-off knowledge, by instead thinking out whether the notion that the observer is present in an on-off manner while consciosness is a persistent property makes sense.

3. A JUST GOOD ENOUGH WORLD MOCK-UP WITHIN THE SKULL

When people ask whether we live in a simulation, they usually mean that some supranatural being wrote and ran a computer program that rules the world we live in. The arguments of the current Section describe precisely a simulation scenario, where the supranatural being is replaced by a functional module of the brain[3], denoted here as diachronic (generative) model, that is continuously devising counterfactual stories.

Perhaps the roots of the modern (hard and easy) problems of consciousness [4] [203, 41] can be traced back to Kant's [154] realization that the noumenon (what it is) and the phenomenon (what it looks like to me) are not the same. It is easy not to notice that we do not directly see the things that lie outside the skull because the brain is a sensationally good simulator of reality.

But although its probabilistic generative model or internal model of the world is accurate enough to fool us most of the time, fooling us is not its reason for being. Its reason for being is informing perception and action selection to further the persistance in time (by e.g. avoiding destructive stimuli or making replicas of itself) of the living system that incorporates it [101], which is a restatement of Darwinian natural selection [54]. A fundamental principle of control theory is that a maximally accurate and simple (good) controller must incorporate an internal model of the controlled system [50]. A living being can be construed as a controller that regulates its environment so as to persist by e.g. avoiding death and making copies of itself, and Darwinian evolution imposes that it must incorporate an accurate enough but "lazy" or having the minimum essentials (simple, lacking superfluous features)[5] controller to do so. Hence, a living being must incorporate a just good enough (generative) model of the world in which it is embedded [50, 101].

Since the internal model of a good controller is a recreation of the surrounding world, the complexity of the model required to persist in time is a reflection of the complexity of the world[6]. The world we live in is in fact by most accounts rather complex, in the sense that it is predictable only to a limited extent: we cannot predict neither too far ahead into the future nor as many diverse phenomena as we would like to, but it is predictable enough that it is worth devoting substantial time and resources in building a model to explain and predict it. This is a consequence of our world being critical [46]: energy is

[3] For the sake of conciseness, in this article we say brain instead of central nervous system (which includes the brain and the spinal cord), which would be more appropriate because the spinal cord incorporates part of the generative model (Section 3.2).

[4] Together with the relevance of the meta-problem of consciousness (see Section 5).

[5] In transaction economics terms, "If we knew enough to be boundedly rational, we would know enough to be completely rational" [157].

[6] In the sense of Kolmogorov or descriptive complexity, which measures in bits the amount of information needed to describe an object.

dissipated in space and time via characteristic fluctuations over a wide range of scales, which entails that many other physical properties also manifest over a wide range of scales in space and time [7], from phase transitions [118] and earthquakes [211] to neural networks [257] and competitive markets [157]. Roughly speaking, a static world in which fluctuations quickly die out without affecting the environment is subcritical. This is an unchanging world, where is no room for either creation or destruction: whatever existed at the beginning, it will stay forever, but no new entities are allowed to join. On the other hand, a supercritical world is governed by ever strongly fluctuating forces that destroy any newcomers that attempt to persist or multiply in it. A subcritical world has low complexity, but creation is not possible in it; a supercritical world both defies predictability and thwarts the maintenance of any particular alien organisms. Moreover, a system poised at criticality may combine versatility and resilience because it can account for complex environmental causes via small parameter changes [194, 133] [7], thus affording brains with optimal computational properties [13]. Finally, a critical world is compatible with new stable structures —which is precisely why we live in a critical world and is a prediction of the anthropic principle [39]— but only on the condition that they are good controllers. They need both senses with a large dynamic range, and a generative model able to accommodate such sensory input and the long-range spatio-temporal correlations of fluctuations. Since correlations exponentially increase model complexity [18], the generative model becomes correspondingly complex, so as to become good enough to regularly foresee noxious stimuli and avoid them. The complex structure of the world is replicated in the multiscale anatomical and functional brain graph properties (e.g. small-worldness and rich-club hubs) which balance information transfer efficiency and metabolic cost [126, 35, 45].

How can one precisely formulate these ideas? Karl Friston's free energy principle[101], which embodies the precepts of Darwinism, optimal control theory [50] and hierarchical Bayesian inference [202], furnishes an information theoretic and probabilistic Bayesian prescription of how any persistent entity or living system behaves. The free energy principle stands on the premise that living systems always try to minimize free energy, which entails minimizing the distance[8] between the approximately estimated (via e.g. a recognition model resting on a mean-field approximation [61]) current probability distribution of internal (e.g. blood glucose concentration) and external (e.g. vision) milieu sensory states and the probability distribution of the generative model (Figure 1), which is determined by both genetic and epigenetic factors or priors and defines the states of the system that are compatible with its existence. Thus, the complex behavior of brains, which day-and-night dynamically and erratically reconfigure their states [292] to keep up with the itinerant dynamics of a complex dissipative environment [98], could be accounted for by solely assuming that the brain goes about its business of descending on its free energy landscape. Its various implementations which differ in how they trade-off accuracy with speed and simplicity[9] [101, 94] meticulously illustrate how complex mul-

[7] This is analogous to how the susceptibility (the variation of order parameters with respect to external sources) is maximal in systems poised at criticality [194, 133].

[8] Kullback-Leibler divergence.

[9] In approximate increasing order of simplicy and decreasing order of accuracy: (free-form) variational filtering and variational Bayes in generalized coordinates, (fixed-form) variational Laplace, DEM, and DEM via iterated conditional mode or ICM-DEM [94]. The combination of variational Bayes, mean-field approximation, and the Laplace approximation is called variational Laplace. For DEM with generalized coordinates schemes, an adjustable approximation is the highest order temporal derivative of the causes that is not assumed to be zero (embedding order) [94].

tilayered hierarchical dynamical models can be fairly accurately and quickly solved and how the required computations can be carried out by cortical architectures [96, 105].

The upshot here is that, while a living agent collects sensations from the world and infers online what are the underlying causes, it assumes that the sensations conform to its *internal* generative model, not to whatever lies outside its brain. This is so because it has no way to verify how correctly its generative model matches reality: it has no choice but to work on the assumption that both its approximate recognition and generative models are good enough. The free energy principle asserts that (living organisms') brains are mostly good enough to ensure survival for some time in the embedding world, but it does not set a specific upper bound to how wrong their depiction of reality can become.

For us, as phenomenal subjects, this means that we are inexorably and unconsciously bound to believe that our inferred guesses about the state of the world are the actual state of the world —which provides an explanation of the illusion of reality (Section 1). In other words, the brain is running a play in which we (d-selves) partake as both onlookers and actors; but unlike in a theatrical piece, we believe it to be reality itself.

3.1. Unconscious perceptual inference versus conscious thinking explicit inference

Inference always proceeds implicitly, but sometimes it is both implicit and explicit. This is plainly portrayed by the contrast between *intuitive physics* and *book-learned physics* knowledge[10].

Intuitive physics is wrought phylogenetically or innately by evolution and ontogenetically by observation and interactive handling of physical objects, and it manifests as the mundane understanding of physics or how objects behave that lay persons possess. It is subserved by the unconscious inferential [129, 121, 101] processes that constitute perception. It is implicit, automatic, and seemingly effortless. It is for instance the intuitive knowledge of the arm strength I need to propel a heavy stone to hit the trunk of a distant elm tree, or how slippery an ice surface can get before I cannot walk on it anymore.

In contrast, reasoning about book-learned physics is explicit, effortful, and it may contradict intuition. It can take over from where intuitive physics offers no answers: it is for instance the knowledge that objects close to the surface of Earth are pulled down by gravity at the same acceleration, or the law of conservation of angular momentum [11]. The modern science of physics is recorded in the corpus of physical facts and theories that humans have gradually accrued along centuries. This information stored in society is accessed by individuals through language and apprehended at an abstract level. Learning it does not build bottom-up from sensory input, but starting from an intermediate abstract level.

Both types of knowledge refer to the same topic, and appertain to the same brain's generative model, but they are different. Our generative model for intuitive physics is, by most measures, superbly accurate for earthly purposes. It is very rare to encounter visual illusions in nature (e.g. negative afterimages due to neural adaptation), but many striking visual illusions have been artificially designed such as motion-induced blindness. The

[10] In general, this holds for any type of knowledge that can be learned under similar dual conditions (e.g. language).

[11] These regularities, albeit elemental theorems for a modern physics textbook, are by no means intuitive: Aristotelian physics [28] did not start routinely coming under criticism until the Middle Ages, and it continued to be studied as physics until the Age of Enlightenment, over almost two millenia.

generative model is *just* good enough: as soon as we leave the realm of daily tasks, and take up say the investigation of microscopic or astronomical events, the incompentence of intuitive physics quickly becomes evident[12]. Then we have to resort to explicit inference at an abstract level, which usually involves scrupulously and hesitantly weaving logical steps[13], not knowing whether we will arrive at the target idea that we want to prove.

3.2. Keeping track of the present: perception and learning

An important aspect of optimizing perceptual inference and action selection is that it is a diachronic endeavor: it requires reckoning with the past, present, and future states that trace an organism's trajectory throughout life. When this information is not fully available —which is virtually always— it has to be inferred. This is a troublesome task.

The world's dynamics at the (macroscopic) scale relevant to human endeavors is characterized by being nonlinear, stochastic, and multi-scaled, with physical causes being stacked in a hierarchy of spatio-temporal nested subsystems. In such environment, exactly inverting the physical causes is in general intractable and often even infeasible [134, 60, 95]. However there exist sensibly accurate and tractable approximate Bayesian algorithms for online inference [5, 18]. Their archetype is the Kalman filter [152], an efficient recursive Bayesian filter (given known system parameters) that alternates between a prediction step (which estimates the present state given the past) and an update step (which uses predictior errors to rectify the trajectory of states veering from the right course) to track the density of hidden causes. The Kalman filter is limited to inference in one-level linear systems[14]; in contrast, natural systems are typically multi-level and nonlinear. This limitation is amended by the extended Kalman filter, which generalizes the Kalman filter to non-linear systems by linearizing the (in general) nonlinear operators with respect to the states, but at the cost of forsaking optimality and reliability[15]. Finally, particle filtering is a family of random sampling-based approaches to nonlinear filtering based on point mass representations of probability densities that makes no assumptions about the form of state-space model operators and noise [5]. Although all these recursive Bayesian filters are computationally efficient, but they are restricted to one level —the parameters and hyperparameters of the system must be known, so the posterior densities of supraordinate causes or states cannot be inferred— and they assume that the noise is an uncorrelated Wiener process [94]. All these limitations are overcome in DEM, a variational Bayesian inference scheme for hierarchical dynamical models in generalized coordinates that is entailed by applying free energy minimization to brain architecture and operation [106, 94, 57]. Not only DEM copes with multiple levels of stacked causes and hidden states [156], but it also handles smooth fluctuations by representing the states in generalised coordinates [94]. DEM proceeds iteratively by updating different groups of internal states and parameters, analogously to the EM algo-

[12]This is also true for something as earthly as the spinning of a heavy object (e.g. a flywheel). We typically do not intuitively understand physical phenomena that are not part of our (or our ancestors') daily routines. Most people cannot tell (intuitively or at all) how a spinning heavy wheel will react to someone tilting it from the wheel plane; even if the person knows the phenomenon from playing with precessing tops or riding a bicycle. Let alone understand gyroscopes or curved mirrors or lenses.

[13]Typically assisted by some mechanical procedure based on logical or mathematical notation.

[14]It is the optimal linear estimator for linear systems consisting of (hidden state) transition and (physical object or cause) observation models with additive white noise.

[15]The unscented Kalman filter was devised to alleviate this unrealibility through deterministic sampling and Gaussian assumptions.

rithm [72, 18] (in which is it inspired). Bayesian inversion in the D-step of DEM operates online like recursive Bayesian filters updating states, whereas the E- and M-steps shape the dynamical operators shapes and precisions at a much slower timescale, as in learning.

3.3. The testing ground for alternative stories: Diachronic, thick temporal models for action selection

In principle, the approximate recursive inference schemes of Section 3.2 could reasonably account for perception, which in essence is the filtering problem of estimating the current posterior density of physical causes given past sensory information. But crucially perception is only useful insofar as it can be exploited to take beneficial action, so perceptual inference and action selection are interlocked in an enactive loop [293, 301]. Perception and action are to inference what the right leg and left leg are to walking. However action selection involves an even harder problem: forecasting or prospecting, which requires estimating future posterior densities given past information[16]. Action selection needs to be proactive and reactive (present and future-focused), as opposed to just reactive (present-focused) to be expedient[17]. The complexity of exact perceptual inference is exponential in the number of hidden causes and states, which is largely determined by the environment; in contrast, the complexity of forecasting grows without bound with respect to the forecasted time span.

In a relentless foray into the future, the brain performs action selection in an exponentially diverging tree of decisions by seemingly nimbly segwaying between nodes or decisions [266, 103, 158]. But if filtering is a difficult problem, optimal forecasting beyond some finite future time horizon is simply unfeasible because the number of possible counterfactual events increases exponentially with time. Fortunately, it is not optimal but good enough forecasting that is required. For instance, the efficacy of heuristic action selection standing on recursive filters has been noted in the extraordinarily complex socioeconomic systems under the guise of disciplined pluralism [18]. Forecasting requires a generative model with diachronic inferential capabilities [19]: inferring temporal sequences of events and to predict the consequences of (and postdict the reason for) one's own actions [302, 31, 104]. The diachronic thickness determines the depth of the tree search into the future of counterfactual stories that has to be expored to yield a solution, and thus the counterfactual richness [249] with its associated computational resources needed to perform action selection. What is the optimal diachronic thickness that balances accuracy and computational resource consumption? Darwinian selectionist imperatives prescribe that how far into the future we actually see is approximately how far into the future we need to see (but not much more, i.e. just far enough[20]) to function and thrive

[16] But this applies equally to any sort of extrapolated counterfactual thinking, such as remembering the past [236], self-projection, and theory of mind [34].

[17] The difference between reactive and proactive is similar to the difference between homeostasis and allostasis. Allostasis proposes that good regulators must anticipate needs and prepare to satisfy them before they arise, whereas homeostasis simply requires satisfying needs as they arise.

[18] "Because the world is complicated and the future uncertain, decision-making in organizations and economic systems is best made through a series of small-scale experiments, frequently reviewed, and in a structure in which success is followed up and failure recognized but not blamed: the mechanisms of disciplined pluralism" [157].

[19] Diachronic models are analogous to thick or deep temporal models [104].

[20] This is analogous to the concept of *satisficing* (satisfying *and* sufficing) in behavioral economics [260] and good-enough comprehension in linguistics [83].

as living organisms[21].

A principled and elegant approach to action selection is minimizing the (expected [22]) free energy of future outcomes [94, 103]. This entails active inference, which is explaining away surprisal through both updating the configuration of internal representations (perception and learning) and acting on the environment [105, 1], in a diachronic manner [158] so that the expected present and future cumulative surprisal is suppressed. This has turned out to be a consistent and remarkably prolific framework that subsumes and accounts for e.g. reinforcement learning [278, 33], optimal control [281] and decision-making [117, 30] theories.

The particular approach used specifies the goal at the computational description level, but the intractability of forecasting implies that the algorithmic level bears the burden of finding a practical solution [188]. Here we hypothesize a solution that is also a fundamental piece that explains the intermittency of the d-self in our model: *short-term forecasting and action selection is carried out deterministically, but long-term involves stochastic (diachronic) sampling*[23]. The reasoning behind this is that short-term forecasting can operate like perception (filtering) while still retaining a tolerable inaccuracy, whereas long-term forecasting precludes finding actions that are optimal enough to be useful via maximization of the future (posterior) density of action (or minimization of free energy) under the system's generative model, because this would entail marginalizing out the blowing up number intermediate temporal variables produced by shifting forward in time the generative model, which typically will lead to a multimodal or nearly flat (and thus unworkable) posterior density of action[24]. The distinction between deterministic short-term and stochastic long-term prospecting has paramount significance for the on-off synecdoche model, as we will expound in Section 3.4 and 4.

Under this hypothesis, easy decisions could forgo short-term forecasting and simply keep the best action (under a given optimization scheme) now as the action of choice for a prolonged time beyond now, or resort to stereotyped stimulus-response mappings[25], such as the muscle stretch reflex (e.g. knee-jerk) and reflexive saccades. Relatively easy decisions could proceed by prospecting the near future and search the action mode (e.g. via DEM) among all the possible prospective future paths. This entails shifting forward the generative model by as much as it is computationally feasible to select the action that minimizes the free energy landscape[26].

Conversely, active inference for decisions involving long-term prospecting (lacking any both more efficient and accurate scheme) would be compelled to resort to a mixture of free energy descent and random hops via sequential stochastic diachronic sampling (similarly to Markov chain Monte Carlo methods) to approximate the posterior density of possible actions. Free energy gradient descent with stochastic perturbations enables

[21] For example, despite functioning in a slower timescale than humans, trees did not evolve to keep a rich counterfactual branching representation of future scenarios because they inhabit a ecological niche where some offspring phenotypes survival is fairly likely by only relying on two exceptionally environmental causes: the diurnal and annual cycles.

[22] Although perhaps one also could make a case for a minimax decision rule.

[23] Diachronic inference, which rests on stochastic sampling, is analogous to the path integral formulation in physics [84] or neuroscience [91].

[24] Variability is injected into the forecasts by stochastic noise terms (which are assumed Gaussian in DEM [94]) in the observation and evolution equations at each hierarchy level.

[25] Direct policy learning, in reinforcement learning jargon.

[26] In theory this is feasible as long as it remains unimodal (i.e. the minimum is identifiable) in the dimensionally much larger parameter space swollen by the new intermediate temporal variables.

surmounting the barriers between local minima thereby improving optimization. The central parameter of diachronic sampling is the temperature, which sets the amplitude of the perturbations. For example, in a simulated annealing scheme [162], the temperature is set high at the beginning and then gradually decreased until e.g. a local minimum is attained. Notably, diachronic inference entails having access to some sort of (pseudo)random number generator[27] in the brain and the production of "fantasies" [134]. The details of the possible stochastic hopping algorithms are not important here, but a plausible approach is variational filtering [91], which samples trajectories online (single pass) via the paths of multiple conserved particles[28] such that the population represents the recognition or ensemble density over states and actions. In principle, this not only allows simulating *single* trajectories arbitrarily far into the future, but also accommodates free-form approximations to the ensemble density of action by forgoing any simplifying assumptions[29] and improves optimization performance by knocking loose processes stranded in local minima [91, 63]. However, as we will see in Section 4.3, in practice it is highly likely that along with of sampling, diachronic inference employs a cut down or simplified version of the generative model in order to further speed up active inference. Finally, the density of states and action represented by the population of particles would be optimized with respect to action. The benefits of diachronic simulations are particularly clear in situations where a deterministic optimizer is stuck in a local minimum (e.g. one cannot find the solution to a problem) and a random perturbation (e.g. quit thinking, rest, and restart, as in a basin-hopping scheme [297]; or unconscious eureka) dislodges the system from the local minimum across a barrier into a deeper ravine (of a new space of possibilities afforded by a new insight).

In brief, the diachronic inferential model is a stochastic and light version of the generative model that enables simulating and optimizing actions for prospective and counterfactual scenarios that would otherwise be intractable to solve (Fig. 1).

3.4. Randomness is invisible

The brain has been under strong evolutionary pressure to compress sensory information as much as possible by isolating the reducible (identifiable or estimable) permanent causal structure from the irreducible noise [8]. This typically leads to representing the presumed environmental states by neural counterparts by e.g. maximizing their mutual information (infomax [180]) or equivalently minimizing the free energy of the (observer-agent) system [101]. Since our generative model has been molded by purely pragmatic (evolutionary) forces, roughly it only has needed to be good enough to ensure that our ancestors had a good chance of survival. Although we cannot know precisely to what degree our generative model —and thus also our subjective experiences— is an accurate account of reality, our very existence is proof that it is fairly accurate (this is possible only because the world that embeds us is underpinned by a constant structured that renders it predictable to some extent).

It plainly follows that any living creature's generative model of its environment can only ever consist of estimable causes or objects —which are defined by the genetical

[27] A plausible candidate for endogenous random generators in the brain is a combination of intrinsic (thermal and fraction of open ion channels fluctuations) neuronal noise [114] and the chaotic itineracy intrinsic to neural networks with balanced excitation-inhibition [290].

[28] Another possibility is the particle filter [5] mentioned in Section 3.2, where point-masses are selected and duplicated or destroyed by resampling through a weighting procedure informed by their likelihood.

[29] For example for variational Laplace: the Gaussian form of ensemble density factors[106]

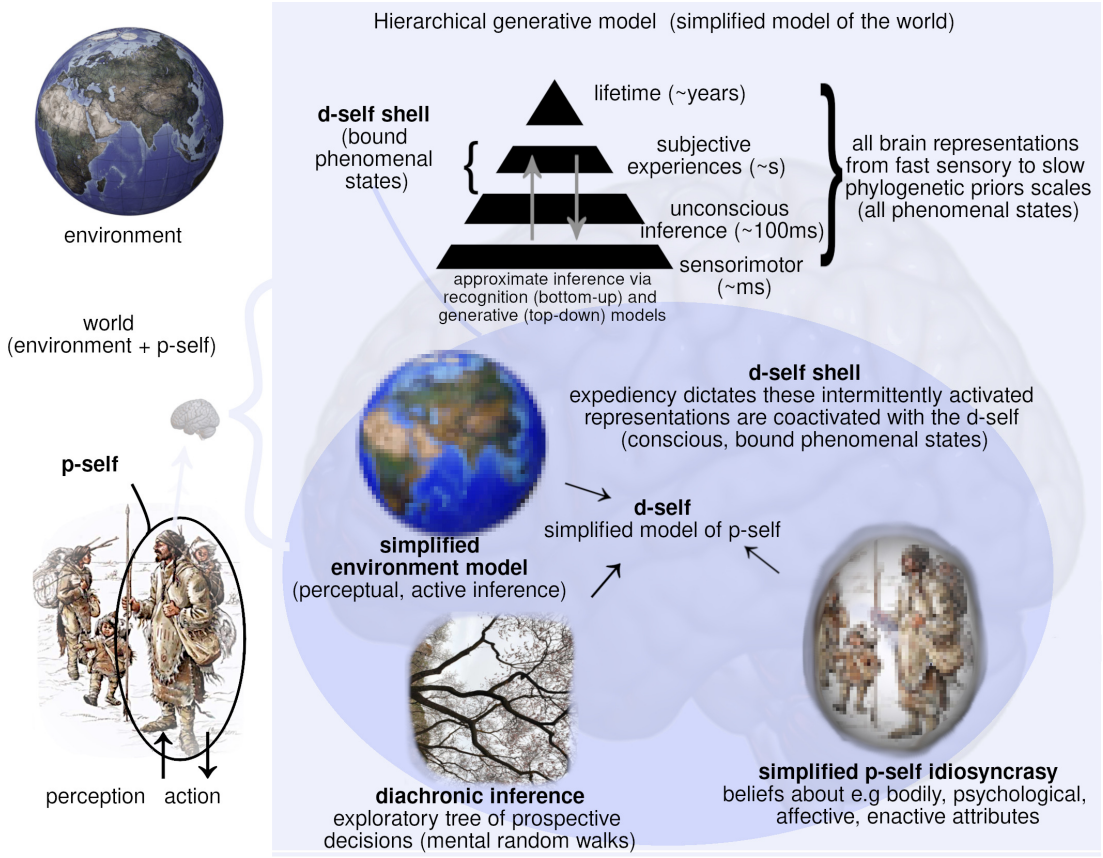

Figure 1: Expediency, approximate inference, and selfness in living systems. A living creature (p-self, on the left) incorporates a simplified generative model (blurry globe) of its world (sharp globe on white background, on the left) that partakes in perception and action selection via approximate inference (through the recognition model) of the state of the environment and forecasting of the consequences of its actions (Section 3.2). The generative model's hierarchy reflects the spatiotemporal scales of the world (pyramid on top). It also enables stochastically sampling trajectories via diachronic inference, which rests on a subset of the whole generative model, typically macroscopic variables that carry the most predictive weight, to simulate alternative scenarios in order to approximate efficient action selection (Section 3.3). Subjective experience at any time is those neural representations (d-self shell, blurry) that are *inferred* attributed or coactivated with the simplified model of the creature or d-self (Section 4.1) because such configuration expedites inference (Section 4.3). Prehistoric hunter-gatherers artwork (bottom left) by Mats Vänehem.

(evolutionary) and developmental history. Hence we are tuned to understand or extract from noisy input only a small fraction of the extant physical objects or causes. For instance, we cannot always "see" or extract a coherent scene from arbitrary visual input: given a digital image of random noise, we are unlikely to find or "see" in it anything beyond a jumble of pixels such as lines, let alone faces[30]. Put simply, given a particular living system, it is more likely to overlook causes or objects to the extent that their causal structure differs from the causal structure of its generative model. In particular, an entropy-maximizing or "shapeless" input is the most likely to be ignored by any creature performing perceptual inference, regardless of its specific generative model: *randomness is "invisible".*

3.5. A hierarchical fabric of loosely and tightly connected processes

The architecture of the brain displays a hierarchical structure that is likely the result of evolutionary pressure to balance energy consumption (efficiency) with representational accuracy [126] in a complex multiscale world. But it is not strictly a directed acyclical graph (or polytree); rather it is a distributed network that contains several intertwined processing streams that encompass many (cortico-basal ganglia-thalamo-cortical [201, 255]) recurrent loops and where links between different hierarchy levels' units are reciprocal (top-down and bottom-up projections) [82, 202], unimodal and multimodal networks have a hierarchical organization, and transmodal networks are assortative [9].

We denote the constituents of the brain graph as nodes, and define them in analogy to the concept of cortical column or module[31] as a domain of the cortical sheet comprising the neurons that encode all possible values of some variable (e.g. an object's orientation or speed or affective value) —which entails that their afferents (and also efferents) roughly link to the same locations.

Here, we assume that the topology of the brain graph governs the degrees of freedom of its constituent nodes such that pairs of cortical nodes are able to function independently to the extent of the (graph) distance that separates them[32] and that its perceptual, attentional, and learning machinery employs unimodal densities [95] as solutions (more on unimodal solutions in Section 4.4). This follows from two interrelated considerations.

First, in order to perform approximate inference, it is efficient to devote computational resources to only track the mode of the recognition model conditional densities instead of attempting to accurately compute the shape of the conditional densities. This can be realized in variational Bayes (e.g. DEM [94]) via the Laplace approximation, which is the assumption that recognition densities take a Gaussian shape (variational Laplace). Assuming a smooth topographical mapping of state values on the cortical sheet —in the spirit of cortical columns and functional segregation [304]— unimodal density functions could be implemented in neural circuits for example as (recurrent) within-level short-range excitatory and long-range inhibitory connections. This would preclude the coexistence of more than one packet of activity within a level consisting few enough nodes. This is manifested for instance as unimodality in the retinal fovea and in early sensory level as center-surround receptive fields [62] and in dynamic field theory as

[30] Despite we tending to go to great lengths to pry out meaningful interpretations, as attested by the phenomena of pareidolia and apophenia.

[31] Which for a given receptive field comprises neurons that encode the full set of values of a feature such as orientation[200, 141, 64].

[32] A tentative more precise formulation would be: their mutual information is inversely proportional to their pairwise (graph) distance.

the Mexican hat-like weight kernels of recurrent networks that accomplish self-sustained (unimodal) activity packets [277, 267].

Second, the brain architecture of bottom-up convergent and top-down divergent connections [93], whereby the quantity of nodes at each level decreases exponentially with level number, entails that the higher the hierarchy level, the less likely it is to accommodate more than one focus of activity (assuming each focus of activity corresponds to a particular inferred solutions or representations). The nodes within a lower level are more likely to operate independently than the nodes within a higher level because most connections are inter-level (top-down and bottom-up), so an arbitrary pair of nodes at a lower level are more likely not to share incoming top-down connections. This is similar to how in an elm tree the paths linking pairs of leaves vary widely depending on whether the leaves belong to the same twig, branch, bough, or only the trunk connects them (Fig. 1). As we will see in Section 4.4, this is essential to understanding (conscious) attention as the effect of a high level node's activity on the rest of the brain graph.

3.6. Four necessary conditions satisfied by subjective experiences

The arguments of Section 3 have outlined an organic framework that characterizes the internal representations entertained by living systems. On the plausible premise that our phenomenal experiences ensue from brain states and processes, it constrains the sort of representations that can possibly be entertained. Crucially, all these conditions hold for both conscious and unconscious representations, but as we will see in Sections 4.3 and 4.4, representations must satisfy more stringent criteria in all conditions to qualify as conscious.

1. Expediency or relevance: Only causes or objects that affect the fitness of the living organism (or its ancestors) will be modeled. Intractable prospective inference may be rendered tractable via stochastic sampling and a suitable simplification in the form of a computationally lighter model (Section 3.3).

2. Estimability: Only modelable and estimable causes or objects can be represented, and thus can belong to conscious experience (Section 3.4). Environmental causes or objects, all else being equal, are "invisible" to the extent of their average descriptive complexity. The estimable objects are encoded in the anatomical and effective connectivity profile of the brain.

3. Unimodality or abductive inference as a compromise between efficiency and accuracy: The topology of the brain graph and its approximate inference scheme entails that representations are predisposed to hold at any one time only one explanation for each inferred object (Section 3.5). This is analogous to abductive inference, or inference to the best (i.e. the mode) prediction [254].

4. Memory intermittency and decay: In tracking environmental causes, the currently activated explanations or internal states continually fluctuate (Section 3.2). This can be seen as an expression of temporal unimodality, akin to a moving or wave packet. Again, to preserve metabolic resources the current explanation will fade to the extent that it is no longer needed for sensible perception and action.

4. Flipping the coin of the consciousness riddle: The reverse is physico-phenomenal states, the obverse is the observer

Now we address the pivotal piece to everting our typical understanding of consciousness: the concept of self (or d-self, see Section 2). But to evert or turn inside out something first we need to know which side is the obverse and which the reverse. Most theories of consciousness assume that conscious experience is an on-off phenomenon switched by some aspect of brain function. Metaphorically, we typically believe that the world is underpinned by a physical structure that we cannot directly experience (the reverse of consciousness) from which our tangible yet intermittent subjective experiences (the obverse of consciousness) somehow ensue.

Here we put forward that a better explanation is: *the world is underpinned by a single physico-phenomenal structure that we can directly experience (the reverse) and our subjective experiences (the obverse) are a subset of it, defined by those brain generative model states that are at any one timepoint linked to the model of self (d-self).* To unravel the meaning of this proposition, it is expedient to make a distinction between bound (in the sense of tied to an observer) and unbound phenomenal states. Roughly, bound phenomenal states are those that an observer knows to be *currently* experiencing [190], typically by virtue of a model of the observer or synecdoche for the p-self (d-self) that ascribes experiences to the observer. A synecdoche is a rhetorical figure where a part of a thing is used to refer to the whole thing. Here we use it loosely as a coarse or simplified model (mock-up) representing the environment in which it is embedded and itself.

What do we really mean by unbound phenomenal experiences, which are not ascribed to any observer? Do they even make sense?[33]. For now it suffices to say that a priori we cannot rule out the existence of phenomenal states that are not attributed to an observer. Whether such lenience is justified or not will be addressed in Section 5. In fact, this dichotomy is very similar to the distinction between non-conscious or conscious and meta-conscious due to Schooler [239], to the distinction due to Block [26] between phenomenal conscious content ("what differs between experiences as of red and green") and access consciousness (contents information about which is broadcast to other brain's systems or "global workspace" [71]), to the distinction between subliminal or preconscious and conscious states [68], between micro-conscious and macro-conscious due to Zeki [305], between lower level access and higher level access [166], and between global broadcasting and self-monitoring [69] (Table 1). Only brain states that can be accessed can involve (e.g. verbally) reports about some aspect or knowledge of the subject's model of itself (d-self) (Table 1, middle column) and thus conceivably qualify as bound phenomenal states.

At the core of these distinctions lies the question of what changes mental states about objects undergo from the first stages of local sensory processing (e.g. the frayed fringe of a cumulus cloud) to latter stages involving concepts in a global environmental level (e.g. the position of the cloud with respect to my body). Psychophysical experiments have shown that in fact we are not able to retrieve experiences that we believe or seem to have experienced (e.g. Sperling's paradigm [269], exclusion paradigm [263], "straining" or overly artificial tasks [239, 192]). This is sometimes explained away by viewers being known to be overconfident about their introspective accuracy [207, 49], which is supported by evidence that categorical (both unconscious and conscious) decisions are systemati-

[33] This appears to be the inverse of the philosophical zombie conceivability conundrum [161], which evokes the image of a "hollow observer" [190].

Table 1: Classification of proposed consciousness terminology into two categories of phenomenal states. Bound phenomenal states require current awareness of the subject experiencing them, whereas unbound states do not (e.g. unconscious states or current guesses about past experiences).

Unbound (always)		Bound (possibly)	Proponent
Non-conscious	Conscious	Meta-conscious	Schooler [239]
Local recurrence		Global recurrence	Lamme [171]
Phenomenal conscious		Access conscious	Block [26]
Subliminal	Preconscious	Conscious	Dehaene [68]
Micro-conscious		Macro-conscious	Zeki [305]
Unconscious	Lower level access	Higher level access	Kouider [166]
Unconscious (C0)	Global broadcasting (C1)	Self-monitoring (C2)	Dehaene et al. [69]

cally and retrospectively biased (perceptual) representations [271, 307, 167, 185]. This is consistent with introspection being an illusion or "reconstructed awareness" in the sense that rather than having insight into our thoughts, we *guess* or infer what are or were our perceptions and thoughts [239, 38] by filling the gaps with top-down priors [166] (Sections 4.3.2 and 6). Here we propose that *bound phenomenal states are simply renditions of past (unbound or bound) phenomenal states recast as in a self-related [47] manner*. For example, thinking of having seen an elm tree yesterday is a bound phenomenal state about recalling, but not about the perception of an elm tree (which only yesterday could possibly have been a bound phenomenal state). Why does it matter that brain states be self-related? The rest of Section 4 is dedicated to flesh out the proposition and to answer the question.

4.1. Guessing what I am: The physical self "out there" or p-self

As any other brain representation, the d-self arose as a means to expedite accurate action-perception. As mentioned in Section 3.3, action can be driven by simple stimulus-reaction reflexes or by the approximate but still onerous solution to inverting the hierarchical generative model projected into the future. The d-self role pertains to the latter process.

A living system (p-self) needs a d-self to distinguish itself from the environment and other p-selves or conspecifics [104], and to understand where it stands and interacts with them. The d-self is a fabricated set of concepts, a collection of related concepts that attempt to just well enough *model* the living system or p-self. Each facet of the d-self is associated with an attribute of the environment that induces the need to model an aspect of the p-self. This is analogous to how, in pragmatic linguistics, egocentric deictic words (symbols or *models*) point to referents (d-self attributes) in a context-dependent manner[34]. For example:

- Space (where?): being here; body ownership and body schema. Usually delimited by a boundary or membrane. Note that non-critical worlds' persistent structures or "creatures" cannot be distinguished from their environment.
- Time (when?): being now, as opposed to memories of the past, or projecting the present self into the past (retrospecting) and future (prospecting; see [236, 34] and Section 3.3).

[34]Namely, "I", "here" and "now" are respectively personal, spatial, and temporal egocentric deictic centers.

- Causality (why? how?): agency, or the state of being the cause of or effecting events, as opposed to the environment being the cause of events [23].
- Other physical variables, such as mechanical force and velocity (how much? whither?): in interacting with the environment, self-knowledge about our physical capabilities and cognitive skills is useful. For instance, knowing our running speed, agility, and strength is useful for e.g. hunting preys, fleeing from predators, and fighting.
- Social and personal identity (who? what?): when multiple other conspecific or similar organisms coexist and coevolve, entertaining a self-other distinction, sensations ownership and empathy, theory of mind, etc. is vital (see Section 4.4.1).

A living system by definition stands out in its embedding milieu, breaking the uniformity of the background. More precisely, a living system breaks the symmetry of all the physical attributes of its environment[35]. It is this "bumpiness" in physical magnitudes that enables incorporating in the generative model a corresponding concept that constitutes each aspect of the d-self and what endows the d-self with a singularly privileged status in the ontology of internal representations. This is useful because insofar as the contemplated action involves at least some part of the agent system or p-self (e.g. its body, goals, or agent-environment interactions, which already covers most cases), a concept of self or d-self becomes necessary to perform action selection. In general, motile entities require much more sophisticated generative models, both of their environment and of themselves (d-self), to be biologically fit: e.g. the "blind and deaf" amoebas require a remarkably sophisticated model of the geometry and texture of their pseudopods to engulf and eat paramecia. The different facets of the d-self are likely to have appeared gradually in the course of evolution, as they became expedient for survival. Just as it is unclear whether transposons or viruses can be regarded as living beings, so the d-self and its attending phenomenological attributes cannot be said to have appeared at a specific timepoint accrued in a piecewise fashion during a protracted geological timescale.

Notoriously there is no conventional definition of what a living system or creature is. But we need at least a working definition because the wide sense of the d-self or subjective self concept refers to a physical counterpart that identifies with "the presumed living system that I am" or p-self. Here we will loosely define a living system in relation to its embedding environment as one that has the ability to preserve and regenerate most of its structure for a long enough time and with a high probability, by spending the (thermodynamic free) energy and materials of its environment. This definition is similar to autopoiesis [291], which further describes a system that can sustain itself by building its own parts. More precisely, an autopoietic system is a network of components and processes (including sensors and effectors), enclosed in a semipermeable boundary or membrane that defines its spatial extent, whose aggregate effect is regenerating and sustaining both the network and the boundary [291]. The notion of living system or autopoietic system is important because it identifies the referent of the self model (d-self) —which is the modeled self (p-self)— with a tangible physical structure.

4.1.1 Motion from emotion

The causes estimated by perceptual inference do not start where the body ends. Inference about external physical objects (exteroception; see Section 3.2) is just one part of percep-

[35]This can be motivated e.g. with synergetics, whereby living systems are construed as persisting patterns that are (at least locally) optimal in dissipating gradients or destroying thermodynamic (not to confuse with variatonal [135, 61]) free energy [123, 286]

tion: inference about internal states associated with both the autonomic (interoception) and somatic (proprioception) nervous systems is likewise imperative for survival.

Inference about environmental and bodily states and action selection cannot be carried out separately. This was envisioned in early accounts of enactive cognition and phenomenology [292], which considered perception as based in the active interdependence of sensation and movement, jointly involving exteroception, interoception and proprioception. Further, the integration of interoceptive and exteroceptive signals within the Bayesian brain framework as being regulated by descending predictions from deep generative models of our internal and external milieu [252] can generalize appraisal theories of emotion[36] by "conceiving subjective feeling states (emotions) as arising from actively-inferred generative (predictive) models of the causes of interoceptive afferents" [248].

This elucidates the processes that transmute visceral and somatic efferents to conscious driving states and emotions. In this view, we can construe driving states and emotions as navigation beacons embedded in diachronic inference. These shape the high-dimensional landscape of the diachronic model so that the action trajectories veer towards particular goals or dispositions (e.g. cold encourages seeking for shelter and anger predisposes to vigorous action). Thus decision making is endowed with a rough bearing (exhaustion, joy, disgust, etc.) that facilitates settling on a course of action —akin to the somatic marker hypothesis [53]— and that we experience consciously as emotions. In other words, driving states and emotions "want" to be acted out.

4.1.2 The embedded, embodied and enacted self

Interoceptive predictive coding models have also been suggested to account for conscious selfness as resulting from embodied cognition: conscious presence (the feeling of sensation ownership) has been tentatively explained away as the prediction error suppression of autonomic and somatic control signals, and conscious agency as being determined by a similary mechanism that in addition accounts for self-generated stimuli [253]. While inference on interoceptive and proprioceptive afferents constitutes the foundation of the embodied self, it is just one facet of the (d- or p-) self.

Although, as described above in Section 4.1, the p-self and d-self are composites of multiple interacting aspects, we can divide them for convenience into three (roughly temporal) categories (Figure 1):

- Perceptual and active properties (short-term, enactive, in a real embedding): body and sensation ownership, (self-)agency, spatial location, body schema (posture), kinesthetic sensations, physiological states, emotions, etc.
- Simulation or diachronic properties (mid-term, enactive, in a fictitious or temporally thick embedding): similar to perceptual properties but in a fictitious or temporally thick [104] context (e.g. ownership of thoughts as opposed to sensations). This occurs during dreaming, action selection, mind-wandering[37], recalling the past, thinking up future plans, etc.
- Idiosyncratic properties (lifespan scale long-term, often embodied): Self-knowledge (what am I like?) encompasses bodily (e.g. hair color, height), psychological (e.g. beliefs, values, dispositions), affective (e.g. temperament) and enactive (e.g. motor

[36] Which assert that emotions result primarily from people cognitively evaluating (as in mulling over) their physiological arousal and external events.

[37] Defined as diverging task-unrelated and stimulus-independent thoughts [262].

skills) properties[38].

The d-self is thus a supraordinate concept overarching all inferred aspects that concern selfhood. It comprises the extended or narrative self, which is the specific history of actions and events or temporal trajectory that defines the present self [205, 112]. Within a hierarchical Bayesian brain, it corresponds to regarding one's body, actions, and its other aspects as the most likely to be "me" [4], which in turn can bring about the illusion of a time-persistent personality core (Section 4.5.4).

However the referent of the d-self —the p-self— does not exist as a clear-cut entity more than e.g. a beech forest: it is a useful concept, but one would be hard-pressed to precisely specify its exact boundary (and say whether the interespersed oaks belong to it or not). Although in Sections and 2 and 4.1 we have attempted to provide a definition of p-self, from the standpoint of the a living being encompassing a d-self, its existence is no more than a belief.

Agency is essential to expedient action selection: it enables distinguishing between endogenous (self-generated) and exogenous events, and rests on prediction of the sensory consequences of movement [303, 301, 110]. This process is called self-specification and consists in integrating efferent and afferent signals to distinguish between reafference (afferent signals caused by the organism's endogenous efferent processes, i.e. limbs) and exafference (afferent signals caused by exogenous events) [294, 21, 47]. Agency is also a major contributor to self-knowledge [112, 140]. Self-knowledge has been tentatively linked predominantly medial areas such as posterior cingulate, precuneus [177, 40], medial prefrontal cortex [238, 58] and inferior parietal cortex [183, 115, 229], whereas the sense of agency has been associated with inferior parietal cortex activity [226, 147]. The neural substrate of many of the d-self aspects seem to be located medially in the brain. Another example is the retrosplenial cortex, which has been suggested to translate between body egocentric and allocentric (other-centered) spatial coordinates, based on its medial location between the hippocampus (which contains place cells encoding allocentric coordinates) and the parietal lobe (egocentric) [186].

Apart from body ownership, the fundamental components of the d-self are the sense of motor agency, thought agency, and ownership of sensations. In particular, thought agency and sensation ownership (and the ability to discriminate between them) correspond roughly to the notion of minimal self [272, 112, 189].

4.1.3 The minimal self

Is it possible to strip away all the unessential[39] features of self and still have an intuition that there is a basic, immediate, or primitive "something" that we are willing to call a self?[112]. Even after dismissing all unessential and conceptual self-referential processes, there seems to remain some sort of conscious selfness [272, 274, 112].

This minimal self (or ecological self [205]) is grounded on a feeling of ownership of sensation and thought that is pre-linguistic, non-conceptual and not contigent on the use of the first person pronoun "I". It is the "self-specifying information attained in perceptual experience[112]. The minimal self is characterized not by being self-evaluative but by being self-specific: it is exclusive (characterizes oneself and no one else) and noncontingent (changing or losing it entails changing or losing the distinction between

[38] Each of these inferred properties is called a self-schema.

[39] Including embodiment and enactivism, according to Strawson [273].

self and non-self) [178, 47]. This conscious aspect of the d-self is likely to spring from self-specifying autonomic and sensorimotor processes [279, 47] that do not require calling on representations of other aspects of the d-self (but see [273]). Hence the minimal self could be defined as (the phenomenal counterpart of) the d-self kernel or self-specific kernel of the d-self. However, it is important to allow for the possibility that *the minimal self may be empty*: after removing all the contingent attributes (d-self shell), we might be left with a phenomenologically hollow entity. Notably, the faceless and immutable nature of the minimal self phenomenologically results in ascribing to oneself a permanent core of identity or I (Section 4.2), which is at odds with the impermanence and multifaceted nature of the p-self.

4.2. Being one self is taking the model of self (d-self) for the modeled self (p-self)

The seemingly prosaic title of this Section states one of the cornerstones of this article. It is likely that after going through Section 1 and having concurred, the reader may still have not suspected that her very sense of self was being called in question —because this might well be the most stubborn of all beliefs. But the illusion of reality also extends to the belief in selfhood. This is because the d-self is just one of the many concepts of the generative model. Another reason is that by construction only a small subset (d-self) of what I refer to as myself (the p-self) is involved in representing itself (p-self), so necessarily the d-self cannot be more than a simplified and lossy depiction of the p-self. The d-self is a just good enough model of the p-self, a homunculus symbolizing a physical creature, a part of the creature modeling the whole creature. In fact, it may be even less accurate than models of external objects the accuracy of these is usually more decisive for survival. Not recognizing that we confound the d-self with the p-self [140] —e.g. thinking that what "I" call I is the p-self— is perhaps the major stumbling block in comprehending consciousness

The term "I" is a first-person pronoun, i.e. a word denoting the person that is uttering it. It serves as the origin point (egocentric personal deictic center) to which subjective experiences are anchored. When I think that I do something[40], I "know" beyond a shadow of doubt that *I* do something, where "I" is not a just a sketch or replica of myself, but *the one and whole* myself in the sense of p-self[41]. But we have established in Sections 3 and 4.1 that I can only have access to a model of myself, i.e. to (aspects of) the d-self, which entails that I *cannot* see (my)self (p-self) directly. Hence it follows that I am confusing the d-self (model) with the d-self (referent).

4.3. Our inner world, especially when expressed consciously, is an expedient and simplified reflection of the outer world

Thinking by itself cannot be expedient: although it may contribute to minimizing free energy, its indirect effect on fitness rests on its modulation of behavior. The notions of agency, ownership, self-knowledge, and minimal self are important, but how do they bear on the distinction between conscious versus unconscious processes? We propose that to

[40] In the sense of I as subject [300, 258].

[41] This is called the immunity principle in philosophy [112]: "When I self-refer I do not go through a cognitive process in which I try to match up first-person experience with some known criterion in order to judge the experience to be my own. My access to myself in first-person experience [...] doesn't involve a perceptual or reflective act of consciousness."

bear the potential to become conscious, an object must be estimable and expedient to the operation of the diachronic model, which works with stochastic sampling on a light version of the generative model (Section 3.3). This corresponds to the first and second conditions in Section 3.6.

My brain clearly knows more than I do: there exist both conscious and unconscious processes. In other words, although every conscious state corresponds to a brain state, not all brain states correspond to conscious states. How can this be explained? Here we propose that, (1) as an offshoot of its normal operational regime, the brain decides or infers *at any specific timepoint* which objects are ascribed to (some aspect of) the d-self or *d-self shell*[42], (2) this decision is informed by a trade-off between accuracy and cost —typically involving stochastic sampling and simplifications (Section 3.6)— that is expedient for the autopoietic system or living creature (fitness or minimizing variational free energy; Sections 3.2 and 3.3), and (3) bound phenomenal states, which are the common sense meaning of conscious experience, correspond to the d-self shell. Here expediency is equivalent to free energy minimization, but perhaps easier to understand as a balance between computational accuracy and metabolic energy saving [191]. The point is that any neural representation consumes resources, so any invoked representation must be beneficial enough to offset its cost (Figure 1). This leads us to: what makes a representation more expedient when it is conscious than when it is not? To answer it, first it is helpful to recast it as: what makes a representation more expedient when it is ascribed to (some aspect of) the d-self than when it is not? That the two are equivalent follows from Section 4.2: activating the d-self *is* what it feels like being one self.

So what makes a representation more expedient when it is ascribed to (some aspect of) the d-self than when it is not? The entity that I call myself —physically grounded on the d-self— is just another neural representation whose reason for being is roughly furthering its host's existence. Crucially, the d-self (and its components) may be an ancillary concept belonging to the diachronic model, which is a cut down, lighter version of the generative model used for complex action selection (Section 3.3). This d-self can be activated and deactivated as the need arises (this is elaborated in Section 4.5). Hence, I experiencing consciously a specific object must be justified in terms of how expedient it is for my host system's fitness. For example, I being constantly conscious of all the subtle fluctuations teeming in my visual periphery could allow me to notice events that I would never know of otherwise, but this is unlikely to be advantageous for my host's fitness (due to being too resource wasteful or distracting from more important goals). This reasoning can be applied to all stages of experience, from perception and short-term memory to emotions and mind-wandering: contents that are more benefitial to the host system in unconscious than conscious form remain so, and vice versa. In other words, the brain incorporates a simulation device (diachronic model) resting on a stochastic and simplified or lightweight physics engine that is tasked with playing out counterfactual stories of the world, where the d-self is one of the many simplified and stochastic simulated objects, and consciously perceived objects are those deemed to be computationally useful in solving inferential problems. Metaphorically, the d-self and its surrounding "shell" of relevant objects will appear as constituting conscious experiences.

[42] Its rough analogy in a Bayesian network would be children.

4.3.1 Free will within nature's Will: A conscious sailboat on a turbulent sea of unconsciousness

Performing action selection requires generating stochastic fluctuations to "sample" possible future outcomes as a way to optimize behavior (Section 3.3). But the the randomness thus infused into our simulations of the world will remain subliminal because it consists of irreducible information and hence not estimable (Section 3.4). This in part explains why knowing oneself is complicated and baffling: since an internal random generator is essential to action optimization, our actions are akin to a boat floating on a haphazardly directed and fluctuating current [91, 63] (Section 3.3).

Free will can be defined as the ability of some entity to make decisions autonomously. Under this definition humans, in the sense of living creatures, are likely to possess free will. However, the belief that *I* possess free will (where "I" refers to some part of the d-self) is false; it is another persuasive illusion[43] caused by mistaking the model of the self or d-self with its referent or p-self (Section 4.2). This is easily understood when noticing that the stochastic and lightweight action selection engine of the diachronic model is *not* part of the d-self: agency —which is an aspect of the d-self— is a higher-order and computationally light representation that is not concerned with how exactly action was selected, but with retrospectively inferring whether the sensory input is caused by the selected action. Another way to look at this is realizing a system cannot be simultaneously autonomous (its decisions are rooted in endogenous causes) and deterministic (its decisions are not random). Hence, if free will exists, it must be (defined as being) unconscious [268] (for neuropsychiatric evidence see Section 4.3.4).

So what is the actual deciding force if not "I"? The answer is the same driving principle of living systems: optimizing action selection to further persistence in time. This driving force can be recast as free energy descent (active inference), perhaps assisted by the diachronic sampler (Section 3.3). Hence, I carry an unconscious engine, guided by free energy minimization, which makes the decisions that (I believe) I make, so in fact I cannot make decisions[44]. In fact, the engine accounts not only for our actions, but also for our emotions and thoughts. Free will is a particular type of causality where the causal effect is attributed to the d-self—where causality is again an internal representation, as envisaged by Kant [154]. This explanation also does away with the ghost in the machine and cuts short Ryleś regression [227], the infinite regress of inner observers that arises when trying to make sense of mind-body dualism. In summary, the belief in free will results from the need of any living creature to work out the causes of its sensations through a simplified representation where the d-self is assigned causal effects (agency)

[43] This persuasiveness was discerningly depicted by Nietzsche [206]: "The thoughtless man thinks that the Will is the only thing that operates, that willing is something simple, manifestly given, underived, and comprehensible in itself. He is convinced that when he does anything, for example, when he delivers a blow, it is he who strikes, and he has struck because he willed to strike. He does not notice anything of a problem therein, but the feeling of willing suffices to him, not only for the acceptance of cause and effect, but also for the belief that he understands their relationship. Of the mechanism of the occurrence and of the manifold subtle operations that must be performed in order that the blow may result, and likewise of the incapacity of the Will in itself to effect even the smallest part of those operations—he knows nothing."

[44] It takes little effort to believe that, from very unalike starting points, Schopenhauer drew a similar conclusion: "Man can do what he wills but he cannot will what he wills" [240]. Schopenhauer argued that no human actions are free, but a necessary consequence of the history and nature of the particular human, similar to how natural causal laws, which he called "Will", govern the world. In brief, the Will is free, but what we call free will is not.

over its behavior.

4.3.2 Mental random walks

Not only action selection but in general any prospective (or more generally counterfactual) mental process is typically brought about by inference resting on (diachronic) random sampling (Section 3.3; Figure 1). Reasoning, intuition, and mind-wandering are kindred processes by which thinking moves from one representation or idea to other related ideas. Although all hinge on the operation of an unconscious and stochastic active inference engine[45] (Section 3.3), they differ in their degree of intentionality or agency: reasoning is akin to a sequence of conscious and intentional mental steps, intuition refers to fleeting sparks of unconscious insight, and mind-wandering alludes to a diffusive sequence of unintended mental steps. All are retrospectively consciously observable, but agency is ascribed to the d-self only for steps of reasoning. In contrast, the causal forces behind intuition and mind-wandering remain unconscious. Other sorts of mental activity such as musing, imagining, pondering or recalling also result in chaining of ideas and recollections that differ in the degree of agency and consciousness ascribed to their underying generator.

Anyway, in general, mental random walks can be construed, similar to a drifting and diffusing particle [94], as being in essence driven by a combination of free energy descent and stochastic diachronic sampling (Section 3.3). By thus exploring the conceptual landscape of possible actions, the living system optimizes its chance of finding expedient courses of action that further its fitness. Crucially, although the algorithm that implements free energy descent and its associated diachronic stochastic sampler remain unconscious (Sections 3.4 and 3.3), the generative model states are potentially accessible to conscious examination when ascribed to the d-self (this follows from Section 4.2). The picture that is starting to come out shows the d-self as an auxiliary function that is only invoked as the need arises (Section 4.5).

Following the beginning of Section 4.3, the driving force of mental random walks is concealed from conscious inspection because the knowledge of its intricacies is unlikely to be relevant to problem solving in ecological conditions. But this clearly gives rise to logical gaps in conscious reasoning, at the higher levels of the generative model hierarchy that include the d-self. How are these gaps filled without conscious knowledge (i.e. without attribution to d-self) of the generating mechanisms? The answer, as with every other hidden cause, is that they are approximately inferred or guessed (see Section 4.3.3). Although this might seem as refusing to look at the evidence to instead concoct it, actually it bears the advantage of reducing the computational burden by recasting low level information-dense representations into high level lightweight representations.

Mental random walks underlie intuition, creative thinking and both unconscious and conscious decisions. Autonoetic cognitive states are those associated with mentally exploring past or counterfactual situations, such as envisioning the future or prospecting, retracing one's past or retrospecting [236], theory of mind and some forms of navigation

[45] This also was presciently described by Nietzsche [206]: "No living being would have been preserved unless the contrary inclination—to affirm rather than suspend judgment, to mistake and fabricate rather than wait, to assent rather than deny, to decide rather than be in the right—had been cultivated with extraordinary assiduity.—The course of logical thought and reasoning in our modern brain corresponds to a process and struggle of impulses, which singly and in themselves are all very illogical and unjust; we experience usually only the result of the struggle, so rapidly and secretly does this primitive mechanism now operate in us."

[34]. These states are subserved by the same medial brain network involved in planning [32, 103], episodic memory [288] and default cognitive states. How can construing autonoetic cognitive states as mental random walks inform our attempts to to understand and improve our reasoning patterns and creativity? For example, the practice of mindfulness [151] can be seen as willfully tethering the (consciously observable part of the) mental random walk with an elastic tether. This theme will be explored in Section 4.6.5.

4.3.3 Filling perceptual gaps onstage: Confabulation in anosognosia

The paradox of being driven (Section 4.2) by something (not what one believes to be oneself) whose (random; Section 3.4) behavioral choices and thoughts one usually ascribes to (what one believes to be) oneself (Section 4.3.1) entails that one must resort to constantly guess in order to explain one's own behavior, and that absolute understanding of the oneself is unfeasible. We have limited access to the causes of our mental processes [207], but we typically do not know it: quite the contrary, by default we tend to believe that we have full access and agency over our decisions (the illusion of free will; Section 4.3.1). This is not strictly a deficit of self-awareness (if by self we mean d-self) but a misattribution of the underlying hidden causes of behavior (Section 4.3.1) grounded on an organic identification between d-self (model of self) and its referent physical object (p-self) (Section 4.2). How do we keep ourselves from noticing that what we call "I" has no saying in decision making? (Section 4.3.1) We confabulate round the clock.

Since our world generative model is remarkably accurate (Sections 3.2 and 3.3), under ecological conditions it is uncommon to come across instances of fabricated explanations. However they are conspicuously illustrated by neuropsychiatric disorders. Agnosia [15] is a non-sensory level impairment of perceptual recognition, i.e. the lack or dysfunction of the neural circuit implementing the (generative model state) representation ("not recognizing something"). It can result from brain injury or stroke, or neuropsychiatric disorders such as post-traumatic stress, schizophrenia, and dementia. For instance, akinetopsia (inability to see motion, often as a stroboscopic effect) and prosopagnosia (inability to consciously recognize familiar faces). Notice that a breakdown of the neural representation itself (e.g. face identification in the fusiform gyrus) entails that both conscious and unconscious processing of its corresponding conscious representation (face recognition) would be abolished. Nonetheless, the breakdown can occur in multiple ways: "The destruction of transmodal epicentres causes global impairments such as multimodal anomia, neglect and amnesia, whereas their selective disconnection from relevant unimodal areas elicits modality-specific impairments such as prosopagnosia, pure word blindness and category-specific anomias" [196]. If a coherent object fails to form (apperceptive agnosia), subordinate level (more elemental) representations may remain intact (e.g. eye recognition) that can make up a makeshift or rudimentary representation that sometimes could pass for genuine recognition (e.g. subtle emotional recognition of a face). Even if a coherent object is formed, it is still possible that identification is impaired (associative agnosia) [14, 15].

How does this bear on the distinction between conscious and unconscious processes mentioned at the start of this Section 4.3? The term agnosia and its offshoots refer to the long-lasting loss of some recognition ability. However, analogous fleeting failures to consciously recognize specific objects can be readily induced by inattention [245] or weak enough stimulation [66, 68]. We advocate that the root of the distinction is the contingent attribution, at any one timepoint, of the relevant state or representation with

Table 2: A double dissociation between visual objects and activation of attribution to d-self.

sensation ownership / sensory processing	deactivated	activated
deactivated	Normal blindness	Anton-Babinski
activated	Blindsight	Normal sight

(a part of) the d-self, where this attribution occurs to the extent that it is expedient (in line with Section 4.3's beginning).

Anosognosia [187] and asomatognosia are types of agnosia in which subjects are unaware of perceptual or motor disabilities or of their ownership of body parts such as limbs respectively; they are deficits of self-awareness ("not recognizing that you do not recognize something"). Anosognosia can cover up *selectively* almost any neurological impairment or agnosia such as hemianopia (blindness in half of the visual field), hemispatial neglect (agnosia of space on one side of the body), receptive aphasia (impaired understanding of language), hemiparesis (paralysis of half body) [289], etc. This hints that in each case of anosognosia the abnormality lies in the neural representation of the *prior expectation* about some aspect of the d-self (e.g. I have two arms, and a right and a left visual hemifield): these expectations are missing or damaged, which leads the subject not to question the absence of e.g. the left side of the body and of the world. Its most remarkable version is perhaps Anton-Babinski syndrome or visual anosognosia [228, 187], a combination of confabulation and anosognosia. It occurs in some (occipital) cortically blind subjects who adamantly affirm that they are capable of seeing while confabulating to fill in their missing visual input. Another case is Korsakoff syndrome [228], which involves a conjunction of amnesia and confabulation.

In confabulation not only subjects do not question their disability or ignorance, but affirm it in the face of opposing evidence. Thus the association between the relevant d-self aspect, which for Anton-Babinski syndrome is visual sensation ownership (Section 4.1.3), and visual input feedback is somehow yielding a false positive despite the total absence of visual input. This fact lays bare the inner workings of brain inference and endorses the notion that our inner world is the result of the continuous and fragile endeavour of guessing the causes of our sensations and the best action to take next. Anton-Babinski syndrome can be seen as the opposite of blindsight, a pathological [231] or induced [165, 176] condition where part of the visual field is not consciously experienced but behavior exhibits some perception. This demonstrates that there can be double dissociations between the conscious belief in perceiving (owning a perception) and the actual activation of perceptual representations (Table 2). Hence, even some of the most fundamental aspects of the d-self, namely the minimal self's ownership of sensations (Section 4.1.3), can be consciously experienced independently from the actual occurrence of perception.

4.3.4 Losing the self: What is looking out from no viewpoint like?

Similarly to the loss of sensation ownership (and its concomitant confabulation; Section 4.3.3), the sense of agency of self-initiated actions (agency awareness) [110, 23] and even thoughts [80, 111] can also be impaired.

Abnormalities in the awareness of action (agency feeling breakdowns) are present

in many neuropyschiatric conditions [23]. For example, damage to the corpus callosum or frontal lobe can result in alien hand syndrome, where the person's non-dominant (usually left) hand can "act up" independently and at cross-purposes with the other hand, which remains under normal volitional control [17]. Here the lack of agency is attributed to a disconnection between primary motor and premotor cortices [23, 6]. Normally, the premotor cortex generates motor commands for the primary motor cortex and an efference copy relayed to the somatosensory cortex that leads to a sensory prediction (corollary discharge; by the generative model) that cancels out the reafference (Section 4.1.2): this gives rise to a sense of agency [294]. But if the primary motor cortex acts up by executing other commands than the ones sent by the premotor cortex, no sense of agency will arise. A subtler abnormality of agency awareness comes about as a positive symptom of schizophrenia (which turn up during psychotic episodes), where patients' actions are consistent with their goals (unlike in alien hand syndrome) but they are not aware of having initiated a movement [109, 23]. This leads to delusions of control (also called passivity experiences), where patients believe their actions to be caused by alien forces. This is typically associated with dissociative experiences: in depersonalization or detachment from the self, patients feel removed from their bodies, sensations, and emotions and feel like their are in "autopilot"; and in derealization, things seem unreal or hazy or surreal and lack emotional coloring, and by anxiety, which exacerbates the feelings of detachment. These symptoms has been explained as the failure of the generative model to produce the corollary discharge that cancels out sensory reafference, while the motor command chain operates normally [23]. Similar symptoms can also be induced via admnistration of dissociative drugs such as salvinorin A, phencyclidine, and ketamine [179, 99].

Confabulation also extends to failures of agency: a fascinating instance is split-brain patients' (whose corpus callosum has been severed) left hemispheres trying to explain behavior that was unknowingly (to the left hemisphere) caused by the right hemisphere [113]. The opposite occurs in the phantom limb syndrome [221], where patients with a recently amputated limb have the feeling that they can move their (non-existing) limb voluntarily: here presumably the belief in agency stems from the predicted state (by the generative model) entering awareness [23] because the current state of the (already non-existing) limb is not estimable anymore.

Schizophrenia deserves particular attention because although self-disorders or ipseity disturbances[46] occur in many conditions (e.g. bipolar disorder and depersonalization disorder) only in schizophrenia severe self-other confusion and erosion of minimal self experience occur [233]. Schizophrenia does not always evolve into psychosis, but its prodromal phase before psychosis is often marked by the basic symptoms [243], which include derealization, disturbances of perception and thought, and exuberance of chaotic and unruly thoughts, which foreshadow the positive symptoms. Although these anomalous self-experiences [214] may be stressful and overwhelming, patients with moderate basic symptoms may be able to keep up normal function; only when the person is no longer able to cope with their symptoms, problems become apparent to others. Crucially, many of the positive symptoms of schizophrenia reflect a failure to integrate intrinsically generated behaviour and concurrent perception [111, 92, 89]. This explains the association of multisensory disintegration with ipseity disturbances via their bidirectional causal

[46]The multiple ways in which disruptions or loss of the minimal self (as opposed to parts of the narrative self; see Sections 4.1.2 and 4.1.3) manifest are generally denominated ipseity disturbances or self-disorders.

relationship: multisensory integration is essential for normal self-experience, and aspects of the d-self such as agency and body ownership are essential to coherent multimodal perception [218, 29, 234]. It has been suggested [218] that hallucinations and delusions may arise from (subconscious) attempts (e.g. hyperreflectivity) to compensate for perceptual incoherence. In particular, first-person perspective becomes dislocated, i.e. the feeling that one observes from an idiosyncratic viewpoint is obscured, and the self-other or self-world distinctions are blurred [232]. The world cannot be apprehend in a holistic or contextually grounded fashion, and the unity of one's own body or thinking are disrupted [29]. Uncontrolled and sundry trains of thought may surge haphazardly and interfere with willed thinking (thought pressure and interference) [130]. This disrupted grip on the world [29] is accompanied by an exaggerated and exhausting form of self-consciousness called hyperreflexivity in which aspects of oneself are not implicitly comprehended but experienced as akin to external objects and are intensely reflected upon in an attempt to gain a grasp on them, and a weakened sense of existing as a vital and self-coinciding source of awareness and action (diminished self-affection) [232, 214, 130].

Strikingly, schizophrenia involves disowning your own thoughts: they may feel as not originating from the self[47] (thought insertion or loss of thought ipseity [214]). This is despite the sense of thought agency being perhaps the most fundamental element of the minimal self (Section 4.1.3). This leads to confusing self-generated thoughts with environmental stimuli such as voices: patients can feel as if their inner experiences are no longer private and may "inspect one's thoughts in order to know what they are thinking, like a person seeing an image, reading a message, or listening closely to someone talking (audible thoughts)" [232]. By analogy with the sensorimotor theory whereby suppression of reafferent sensations brings about the feeling of agency (Sections 4.1.2 and 4.3.3), it has been proposed that thought agency arises when "thought corollary discharges" predict actual thoughts [80, 111]. This is supported by patients with thought insertion (thought misattribution) also displaying delusions of control (agency misattribution) [111, 112], if we conjecture the existence of a faulty prediction error suppression mechanism engaged in both sensorimotor reafference and "thought reafference".

But how exactly is (endogenous) thought discriminated from (exogenous) perception or (endogenous) dreams? We propose a scheme, the "mischievous random walk", that hinges on the concurrence of unsuppressed thought reafference and the mental random walk analogy (Section 4.3.2). In principle, for perception it suffices to take the result of inference on any (exogenous) sensory input at face value and consider it as one's own perception. But thoughts are (1) endogenous processes: they are different from either endogenous motor commands matched with reafference of agency or the pure exafference of perception. Thoughts are also (2) chiefly top-down processes: perception is mostly driven by bottom-up prediction errors, whereas agency is evenly reliant on both top-down predictions and bottom-up afference. But even more distinctively, we hypothesize that thoughts are (3) trajectories that unfold top-down by stochastic sampling of a generative model seeded with a suitable configuration (e.g. what is likely to happen if I propel upwards the stone that I am holding in my right hand?). Since the transition function of my thought steps (the multiple ways in which different ideas transmute into each other) is probabilistic, a thought started with the same seed will yield (perhaps slightly) different trajectories every time. These characteristics suggest a simple algorithm to discriminate thoughts from percepts: checking —typically at a middle-high cognitive

[47] Curiously, this seems to refute the immunity principle [112] (Section 4.2).

Table 3: Self-disorder anomalous experiences (from the "Cognition and stream of consciousness" domain items of the Examination of Anomalous Self-Experience questionnaire [214]) as anomalous mental random walks. The failure to suppress thought reafference leads to the loss of the sense of thought agency, so chains of thought (generated by random diachronic sampling) are perceived passively as "mischievous" or unwilled mental random walks.

Anomalous self-experience	Mental random walk	Algorithmic level cause
Thought interference or popping up	Unchecked mental random walks that take away computational resources	Spontaneous initiation of diachronic sampling, mismatched thought reafference
Loss of thought ipseity	Disowned chain of thought	Mismatched thought reafference
Thought pressure or overflow	Random sampling of parallel chains of thought	Failure to prioritize one single chain of thought (Section 4.4)
Thought block	The chain of thought stops	Sampling halts spontaneously
Thought perseveration	Unwilled prolongation of the current chain of thought	Failure to extert control on the sampling process
Perceptualization of thought (e.g. thoughts occupying physical space or being heard by others), confusing perception with imagination	Self-generated chains of thought blended with exogenous sensations	Mismatched thought reafference along with unsuppressed sensory input
Confusing memories with dreams	The chain of thought is perceived as an unconsciously generated dream	Mismatched thought reafference
Discontinuous awareness of own action Subjective time speeding up, slowing down, standing still, becoming fragmented	Chain of thought broken into pieces and reassembled	Abnormal attribution of thought chains with the d-self (Sections 4.2 and 4.4) and/or faulty retrospective reconstruction of thought chains, cf. [189]

hierarchy level [159, 191]— the conjunction of (conext-seeded) top-down activity streams triggered by a (pseudo)random generator (Section 3.3) with the absence of bottom-up input. Remarkably, many of the cognitive self-disorders characteristic of schizophrenia can be parsimoniously explained through a combination of mental random walks with failures to match thought reafference, which result in alienated or "mischievous" random walks that are experienced as a passive observer, as opposed to as a thinking agent (Table 3).

Astoundingly, schizophrenia seems to be capable of disrupting *all* the neural mechanisms that specify each aspect of the (p-)self-other distinction (Section 4.1) and hence the d-self, up to the minimal self [189] (Section 4.1.3). It can induce symptoms such as depersonalization, blurring of demarcation ("person confuses their thoughts, feelings, and other aspects with their interlocutor, or otherwise feels invaded or intruded upon by their interlocutor" and "confuses themselves with their reflection, such as when looking in a mirror" [214]), d-self disintegration ("person feels as if their experiences aren't their own, at least briefly, or as if they were a mere inanimate object" and "person feels an incredible distance between the self and experience, resulting in intense and involuntary constant or recurring self-monitoring" from the Examination of Anomalous Self-Experience or EASE questionnaire [214]), and splitting the subjective experience of the d-self, e.g. "sense that the self does not exist as a unified whole beyond having a multifaceted personality (I-split)" [214]. Crucially, schizophrenia can also induce a dimished transparency of consciousness ("a sense that one is blocked from clearly perceiving the contents of con-

sciousness", from EASE [214]), which under our scheme we can explain as an abnomal attribution of experiences with the d-self (Section 4.3). Similarly, diminished presence ("increasing distance from the world experienced as apathy towards specific events" [214]), diminished initiative ("pervasive difficulty initiating goal-directed activity" [214]) and hypohedonia can be construed as abnormal coupling between the d-self and the free-energy landscape navigation beacons or emotions (Section 4.1.1).

But what does it mean, phenomenally, to perceive the world or think while the d-self falls apart? Can an observer be hollow [190]? Someone who has never had psychotic or dissociative episodes may not be able to imagine what it is like, but accounts of schizophrenia patients deprived of "the solid center from which one experiences reality" [230] suggest it might be possible. Persons suffering schizophrenia exhibit dialipsis: the disintegration of the d-self tends to fluctuate over time based on emotions and motivation [233]. This affords a singular look into how a disintegrating (or reintegrating) observer might experience consciousness, as the d-self falls apart or its attribution with internal representations or shell is disrupted —and phenomenal experiences shift back and forth between being bounded or unbounded.

The neural basis of schizophrenia is not well understood. The dysconnection hypothesis states that schizophrenia stems from abnormal interactions between different areas, not only at the levels of physiology and functional anatomy, but also of cognitive and sensorimotor functioning [92]. These are caused by abnormal neuromodulation of synaptic efficacy in specific brain systems, particularly prefrontal, which result in the "inability to augment (attend) or attenuate (ignore) the precision of sensory evidence, relative to the precision of beliefs about the causes of sensory cues" [2, 99]. This in turn compromises learning that rests on activity-dependent associative plasticity, which could explain reduced hierarchical multimodal network organization and the loss of frontal and the emergence of nonfrontal hubs in schizophrenia patients [9, 90].

The takeaway from the anomalies in the awareness of action and thought agency and its sometimes attending confabulations is firstly that all (conscious or not) aspects of constituting the d-self (Section 4.1) —even the most fundamental such as body ownership and agency (Section 4.1.2) and ownership of thoughts (Section 4.1.3)— are susceptible to disruption. Secondly, these awareness deficits have been accounted for by positing that while predictions are accurate (reafference suppresses corollary discharge) only certain components of the world's internal representation are available to awareness, viz. affordances and desired, predicted and estimated actual states, whereas motor commands, and actual movement and states and sensory feedback remain unconscious as long as the discrepancy between the desired and reached states is not too large [23]. As we will see later (Section 4.5), this is consistent with the proposal that access to consciousness is determined by attribution to the d-self (Section 4.3). Briefly, this is because the d-self is a higher order construct that is activated precisely to solve such discrepancies.

4.4. Why am I just one I?

Here we expand on the unimodality of brain internal states' conditional densities (the third item of Section 3.6) in relation to the d-self. An unimodal[48] probability distribution is manifested as an unequivocal belief in one option, as opposed to an ambiguous belief that assigns similar odds to multiple options. This is closely related to attention,

[48] In the statistical sense of a distribution with a single mode. Not to confuse with unimodal association areas, which are brain areas that process information of a single sensory mode.

whose foremost role is transforming "ambiguous" multimodal, uninformative or flat (high entropy) prior distributions into peaked unimodal (low entropy) conditional or posterior distributions.

Attention is typically understood as the allocation of limited cognitive processing resources [3]. In performing both perceptual and active inference, the brain needs to throw out some —typically all except the maximum— low probability explanations to keep the computations required for inference light enough to enable real-time reactivity. This is congruent with the states available to (conscious) introspection (following selectionist principles) being constrained by expediency and simplicity (Section 3.3 and 4.3), e.g. by being required by the diachronic model to efficiently compute the next optimal action given the currently estimated current world state. This informational pithiness is reflected in the (presumed) approximate inference schemes employed by the brain. This is notably straightforward in variatonal Laplace [106] (a fixed-form variational Bayes inferential scheme, see Section 3.3), which enforces the unimodality of recognitions densities during (abductive) perception and action. In this scheme, bottom-up and top-down attention can be construed as enforcing strong enough unimodal distributions on the recognition density and the density of the currently sampled diachronic trajectory or simulation scenario.

Hierarchical inference involves a delicate balance between bottom-up driving sensory error signals and top-down predictions. This is accomplished via the saliency or precision parameters that modulate the gain of driving signals. It has been suggested that attention can be understood as inferring the level of uncertainty or precision during hierarchical perception [81], which roughly corresponds to the (optimized) width of the unimodal distributions. The richness of configurations afforded by this scheme could mechanistically explain e.g. the pathologically narrow attentional focus domain found in simultanagnosia [168] or how attentional resources are flexibly allocated depending on task demands [296]. Exteroceptive and interoceptive gain modulation would correspond to top-down attention being focused on environmental and bodily sensations, respectively.

The roughly tree-like hierarchical architecture of the brain (Sections 4.3.3 and 3.5) goes along with its propensity to render unimodal the conditional densities that it entertains. Higher-level or slower cognitive concepts lie near the narrow top or "trunk", whereas lower-level of fast sensory features lie near the wide bottom or "leaves" and "twigs" (Figure 1). This topology enforces a conciseness and unity in the array of simultaneously held higher-level interpretations about the world at the "trunk" because the higher a concept or explanation lies in the hierarchy, the more likely it is to be laterally connected to all other homologous explanations (as in a winner-takes-all scheme [224]). In intermediate levels, a few laterally unconnected "boughs" could support the expression of uncertainty as divided attention [122]. In the lower levels, many "branches" or "twigs" can function with a high degree of independency as split processes, when they are neither connected laterally nor governed by common top-down predictions. Typically, any pair of branches can communicate via their first common parent branch, but the communication strength decreases with the number of interposing nodes linking them. Overall, a hierarchical inference system incorporating strong enough informative priors tapering off toward the top has an organic predisposition to settle at any one timepoint into a single higher-order explanation —e.g. world interpretation, prospecting simulation, or reminiscence.

We suggest that, analogously to how simplicity-inspired unimodality or abductive

inference explains that the brain leans toward one single perceptual explanation[49] (we typically see just one scene at any one timepoint), unimodality explains the unity of the d-self and the oneness of the focus of attention during thought, action planning, daydreaming, and in general autonoetic consciousness. A compelling case for unimodality or abductive reasoning is that action selection unavoidably must be followed by action execution —which by definition must be just one— so any choice symmetry between multiple courses of action must be ultimately broken. In brief, having one body translates into having one main focus of attention [76]. Finally, unimodality was also recycled and redeployed for diachronic modeling purposes.

How is all this related to consciousness? Attentional mechanisms resting on gain modulation circumscribe the type and amount of plausible solutions (for both perceptual and active inference) that are entertained at any one time. This together with the constraints inherent to variational Laplace (discussed above in this Section) determine to a large extent the inferential dynamics of the brain. Hence, expediency or free energy minimization and computational efficiency considerations (Sections 3.3 and 4.3) can account for the definiteness (oneness) and informational content of conscious experience. However, the attentional reshaping of internal states per se does not determine the contents of consciousness [171, 164]. To ascertain which "branches" and "twigs" of the tree of potential conscious experiences, attribution to the d-self must be established (Section 4.3).

4.4.1 The others like I

We are not alone in this world; we humans tend to live in groups. This induces the need to explicitly model who I am to enable the distinction between "I" and "they" (Section 4.1) not just as distinct lumps of matter, but also as distinct agents endowed with generative models that are analogous to mine. Because we are fairly complex creatures, including the states of conspecifics in the set of world states vastly exacerbates the problem of inferring world states[50].

Can a creature that lacks theory of mind entertain the concept of being consciously aware of itself? [104] A creature that lacks d-self cannot even entertain the concept of itself, but a creature without theory of mind could still possess other useful aspects of d-self such as bodily and minimal self representations (Section 4.3), depending on how theory of mind is implemented. This question is an expression of the meta-problem of consciousness [43], which in turn is tightly intertwined with our particular beliefs about the concealed machinery of the world, or physics (Section 5).

[49] On the face of it, this may seem to bear on the von Neumann-Wigner interpretation of quantum mechanics [295]. In quantum mechanics, the wavefunction describing a system tends to spread out into an ever-larger superposition of different possible situations until an observation collapses it into just one outcome. Roughly, the Neumann-Wigner interpretation puts forward that human observers do not sense superpositions of multiple outcomes, but only one outcome, despite a superposition of observers seeing different things being the natural state of affairs before the wavefunction collapse, so it must be that consciousness causes this collapse. However, our discourse is agnostic about how wavefunction collapse occurs. The unimodality of conditional densities is entailed by only computational efficiency arguments. At no point we bring into play quantum effects to explain brain functioning —not even to account for the stochastic sampling of the diachronic model in Section 3.3. For example, the many-worlds interpretation [79] can also explain the appearance of wavefunction collapse through quantum decoherence without bringing up consciousness.

[50] Although sometimes elegant simplifying assumptions can be applied, such as assuming a permutation symmetry between me and other conspecifics, i.e. that we all possess the same model as I do [38, 100].

4.5. The on-off self

Here we explore the implications of having intermittent neural representations (fourth item of Section 3.6) of the d-self. In Section 4.3 we causally stated that conscious experience occurs when a representation is ascribed to (some aspect of) the d-self (i.e. it is determined by the d-self shell). Although this follows from Section 4.2, it deserves further elaboration.

4.5.1 You (or your d-self) have been unknowingly hired as a middle manager

In life, one is thrown into an unknown world that one must progressively learn to understand in order to achieve some preset goals. These goals are the generative model's top priors which, having been handed down through many generations of our ancestors, contain valuable parameters that expedite survival [101]. The top priors are likely to sit in subcortical regions such as the hypothalamus, which is a key regulator of hormones secretion and overseer of vital drives such as hunger, body temperature, maternal attachment, and fear. Not by chance, these are among the most powerful driving states that determine our behavior and thoughts: there is only so much that I (my d-self) can do while disregarding the "directives from the top". This is translated phenomenologically as an uncomfortable and excitatory state caused by a physiological need (e.g. thirst) and an urge to suppress it (finding water) [143]. In other words, the director or senior managers of the brain hierarchy govern the wishes and expectations of the d-self: we are enslaved ("enslefd") by means of being told what we want (again, "Man can do what he wills but he cannot will what he wills" [240]). Crucially, although we are told what the goal is (quenching thirst), we are not told how to achieve it. Our job and purpose in the game of life is precisely figuring this out —but of course, as explained in Section 4.3.1, the belief that "I decide" is also an illusion (of free will).

Some problems can be solved without the participation of a model of one's own body or agency, such as thinking of why the sun keeps regularly turning around. But others require to model some aspect of the d-self, such as hand-tool coordination and social communication. This is where we step in, but not always. Most of the (bottom-up) error signals are suppressed soon after arising in the lower levels of the hierarchical generative model. This is evinced by the brain's workforce continually making fine adjustments (to improve perceptual inference and action selection) which are unavailable to awareness [101]. Crucially, only when the discrepancy between the sensation and its prediction or the intended and actual movement are large, the error enters awareness [279, 128, 23, 22], i.e. it is ascribed to the d-self. From our (d-self) perspective, this is felt as slowing down and switching from autopilot to conscious thinking. Once a problem is thus conveyed to the middle manager or d-self, its task typically involves working out unexpected situations or solving novel problems, within a vast space of possibilities, in a simplified model of the world. This is precisely because the d-self is a lightweight device with little turnover (Section 4.3): it (and thus "I") would be overloaded with information if it ("I") were aware of any discrepancy and fine adjustment elicited by small ascending prediction errors. In fact, the focus of attention is so narrow that we overlook most of the small environmental fluctuations (decoupling) [261]. This limitation of computational resources is a way to negotiate the balance between metabolic energy sparing and computational efficiency [126, 191], which ultimately determines the contents of conscious experience.

So although we cannot avoid feeling in control and possession of our body and actions, in reality we are just moderately important (but still dispensable) subordinates in the brain hierarchy of command, and we (d-selves) are sporadically invoked only when needed.

4.5.2 Turning on, refreshing, and turning off the d-self

We are inclined to believe that we exist as continuous entities or selves through time. This is also an illusion, just as the illusion of reality (Section 1) and of self (Section 4.2). This follows from the transitory and fickle quality of all neural representations, which are turned on and off at the brain's discretion, in attempting to further its existence (or minimize free energy).

The d-self is a stand-in for the physical embodiment of the self (p-self) in our internal simulation of the world. As such, the d-self is a computational device that is invoked only when needed (Section 4.3). Its intermittency is likely to ultimately be an expression of the mostly unpredictable phenomenon of on-off intermittency [217], induced by perceptual self-organized instability [98] and is likely associated with the intermittent up-down states of membrane potentials in neurons [191]. Therefore, our subjective experiences ought to be intermittent in an similarly unpredictable manner. In fact, we spend most of the time unconsciously [198]. But clearly the p-self persists even if its neural representation or d-self is being intermittently activated.

The temporal lapses in the history of d-self activity are translated phenomenologically as I often trying to recall or infer what I just did when "I (in the d-self sense) did not exist". The confusion between the p-self and the d-self lies at the core of the puzzle of consciousness. This has also been expounded as temporal dissociations, that occur "when an individual, who previously lacked meta-consciousness (intermittent explicit re-representation of the contents of consciousness) about the contents of consciousness, directs meta-consciousness towards those contents; for example, catching one's mind wandering during reading" [239] (cf. Table 1).

The implications of the d-self being an intermittent entity for the study of consciousness can be illustrated with the following metaphor. The usual attitude toward consciousness is that of an observer's attribute that exists intermittently, corresponding to alternating conscious and unconscious timespans. We argue that this view is misleading. It is more illuminating to see consciousness as a screen or television that is always on. Instead, what is being turned on and off is the d-self or observer, as is the case with any other brain representation. Hence, what appears as unconsciousness is actually selflessness.

4.5.3 Time-discrete d-self packets as sesmets

Although physical time is usually regarded as a (spacetime) coordinate or a parameter of a dynamical system, in fact it is unknown whether time is continuous or discrete. If time were a discrete process[51] or a process with a fundamental period below a specific upper bound of $\approx 10^{-33}$s [298] (which is much longer than the theoretically smallest observable timespan or Planck time), we would not be able to tell the difference from just observing physical systems.

[51] This suggests an alternative, more speculative account of how objects are attributed to the d-self (Section 4.3). Perhaps neural representations are attributed to the d-self simply by being coactivated with the d-self in the same (discrete) time slice.

Even if time is continuous, the operation of the brain relies on all-or-none, discrete events —action potentials or spikes— to transmit signals between its computational units or neurons. Further, neurons cannot fire continuously: they have a refractory period after before which they must remain quiescent after firing a spike. Hence neural computations are essentially discrete-time processes[52]. Thus we can regard any physical object, including the brain and representations such as the d-self, as a series of static snapshots.

Galen Strawson sought the most stripped-down version of a self that can still be called self, and was led to define the (synchronic) self, *sesmet* (subject of experiece that is a single mental thing) or sestem (sentient system over the time scale that it persists [108]) as a subject of experience that is a single (discrete) mental process during a hiatus-free period of thought [274]. Each sesmet is short-lived and has no history, and consciousness consists of a sequence of sesmets. The sesmet can be portrayed as a temporally finite "packet" of d-self activation, with its shell. Importantly, this entails that sesmets determine the temporal extent and content of bound phenomenal states (Section 4 and 4.3). It lasts just long enough to accomplish the task that it was born to deal with, and then it fades away. The sesmet is the atomic unit of conscious self-awareness[53]. This formulation is consistent with the notion that the d-self is invoked only if prediction errors are encountered [279, 23] as with conscious events or sesmets being discontinuous, punctuated, and possibly incongruous with each other (across time).

Here we view a sesmet as each of the non-overlapping episodes or instantiations of d-self (including minimal self) activation, upkeep, and deactivation[54]. Hence, we (as d-self instantiations or sesmets) exist only as snapshots including a d-self interlaced in a daisy chain of snaphsots that are updated at discrete intervals. Each snapshot can contain not only present, but also past and alternative future events[55]. These various components require a frame or window of simultaneity that corresponds to the duration of lived present[56] [292] (cf. diachronic thickness). The duration of a sesmet is related to the thickness of the diachronic model (Section 3.3). Typically, it is a few hundred milliseconds [159, 237, 132, 191], which appropriately corresponds to the timescale of the fastest macroscopic (our body size) fluctuations in our environment (Figure 1). Hence sesmets are not updated faster than this because typically it is not necessary to keep up with world events.

Each sesmet is self-contained and independent, and may comprise a diachronic snapshot of the past, present[57], future, or in general any other counterfactual timeline (akin to Husserl's three-part structure of temporality [292]). As usual (Sections 3.6 and 4.3) its

[52]This is similar to how the components of a microprocessor operate at clock rate periods, but importantly without being yoked by a central clock generator.

[53]Curiously, this implies that Boltzmann brains can be replaced by "Boltzmann sesmets". Boltzmann brains are full-fledged brains with our typical thoughts that theoretically could spontaneously pop up in void as a result of quantum foam random fluctuactions in an empty universe. They are a thought experiment used in physical cosmology to set a lower bound for how unlikely is our university to exist.

[54]To avoid concomitance of multiple sesmets, we assume that the minimal self is unique by construction.

[55]Or retention, present, protention, which constitute the three-part structure of temporality gleaned by Husserl using phenomenological reduction [292].

[56]"In this view, the constant stream of sensory activation and motor consequence is incorporated within the framework of an endogenous dynamics (not an informational-computational one), which gives it its depth" [292].

[57]Strictly speaking, the present is actually part of the future and thus unknown because both sensory and hiearchical inferential signals propagate at a limited speed.

contents are determined by an array of factors evaluating their contribution to the host system's fitness. Thus, for example, sesmet contents are *inferred* to have been perceived by or affected the d-self in a pre-sesmet time period. Notably, the events or objects that have been consciously experienced (i.e. included in past sesmets) leave a stronger memory trace and hence are more likely to be picked up by subsequent sesmets. By limiting the size of the buffer of inference-relevant items, diachronic thickness —as an expression of short-term memory decay (fourth item of Section 3.6)— plays a major role in our experience of consciousness. From our (d-self) perspective, the computations redrawing the "canvas" of consciousness proceed automatically and unconsciously [132], and the "canvas" corresponds precisely to sesmets or the snapshots that include a d-self. In this sense, in fact we are turned on and off not just with a period of one day (sleep-wake cycle, Section 4.6.3), but also with subperiods on a subsecond scale.

4.5.4 The stream of consciousness and the narrative self as nested, not interlaced, sesmets

"On the one hand, there is the present as a unity, [...] our abode in basic consciousness, and on the other hand, this moment of consciousness is inseparable from a flow, a stream" [149]. In the coexistence of permanence and change, consciousness is a constant background against which distinct temporal acts and events appear[58] [292]. What is the nature of this sense of a continuous self? Is it carried by a succession of momentary minimal selves that are tied together by real connections? [112] In other words, what is the relationship between the sesmets and the narrative self (Section 4.1.2)? Hume suggested that the (narrative) self consists of a bundle of momentary impressions (sesmets) that are strung together by the imagination [144, 112]. A similar idea is found in buddhist texts, which consider the stream of consciousness as a "skandha" or aggregate of composite entities without independent existence in the form of "a series of rapidly changing interconnected discrete acts of cognizance" [27] or "a sequence of momentary mental states, each distinct and discrete, their connections with one another being causal" [155].

But perhaps a better answer lies in the concept of generalized coordinates, which are explicit (neural) representations of temporal derivatives of physical variables [94] (Section 3.2). For example, most of our conscious experience of velocity does not stem from incremental changes of position, but as neural representations of velocity overlaid on images [101]. In other words, motion is "painted" on top of static pictures. This remarkable mechanism is laid bare in motion aftereffect (where motion is perceived in the absence of motion, due to motion adaptation) and the visual disorder of akinetopsia (where motion perception is lacking despite observing objects being displaced). The brain activates sesmets that include static pictures with generalized coordinates[59] such as velocity that contain information about the temporal evolution of objects, which are consciously perceived as motion.

Although sesmets occur as discrete sequential disjoint events, each sesmet is a functional independent unit containing the information relevant to optimize perceptual and active inferece, including some information about past sesmets and prospective future sesmets. Once the sesmet is constituted, it is virtually causally independent from all

[58] This was called "double intentionality" by Husserl, since there is not only a retention (of the object event) but also a retention of retention (a reflective awareness of that experience) [292].

[59] This is similar to how a discrete sequence can be equivalently represented with the information contained in its derivatives at one timepoint, using a Taylor expansion [94].

other sesmets (especially future ones [131]). Past sesmets have existed embedded in a time sequence, but from our subjective (d-self) perspective, they are recast synchronically into the present sesmet as a static and simplified representation of the past, i.e. their useful information is nested into the current sesmet, which carries temporal information both in the form of time-event pairs and of temporal derivatives. To sum up, sesmets manifest as a sequence of on-off brain activations, where each sesmet comprises an approximate static representation of the past, present, and future of the sequence in which it is embedded; we experience these static snapshots as the (dynamic) stream of consciousness.

4.6. Inferring the inferrer: from interoception to autoception

The generative model of the brain is capable of inferring environmental states of the external world (Sections 3.2 and 3.3) and bodily states, including homeostatic and visceral variables felt as drives and emotions (Section 4.1.1). These exogenous states induce afferent signals relayed into the brain by sensory sheets and homeostatic receptors. But what happens to error signals once they enter the hierarchical inference engine? For example, is the reaction to input of visual area 1's neurons or our own appraisal of our emotions also inferred? Do we (our brains) try to predict our next thought? In other words, does the generative model also infer its own signals?

4.6.1 A finite stack of cortical sieves winnowing sensory input

The generative model is like an imperfect mirror of its surroundings. Does it need to create a reflection of a reflection? The hierarchical structure of the brain mimics and emerges from the separation of temporal and spatial scales of its environment [123, 191]. In perceptual inference, the same algorithm seems to run at every cortical stratum of the brain scaffoling [95, 141]. This algorithm is invariably some implementation of infomax [180], i.e. maximizing the mutual information between incoming and outgoing signals, which entails wringing out as much redundancy as possible. Arrays of signals are sequentially compressed into increasingly lighter representations: each supraordinate level represents a dynamical structure whose complexity is less than that of its subordinate homologue [96], while the irreducible stochastic noise is left unmodeled [191] (Section 3.4). The hypothalamus and hippocampus are likely to sit near the apex, establishing respectively the top genetic and epigenetic priors. Under this scheme, each level is already trying to predict or suppress the ascending input or prediction errors from subordinate levels, so another separate modeling subsystem for the model (a synecdoche of a synecdoche) seems to be redundant. There are however two exceptions: internally generated noise (next Section 4.6.2) and the generative model hierarchy top priors.

The plain reason is that the internally generated noise through diachronic inference and its ensuing actions and the top priors are the only subsystems of the brain generative model that contribute new complexity to the world[60] in the sense that the rest of the generative model is trying to mimic the world, including both the outside and the inside of the body, *but not the generative model itself.*

The top priors define idiosyncratic properties of the living system (Section 4.1.2) that typically characterize it and its ecological niche for most of its lifespan, and are largely

[60]This is only true to the extent that the generative model is a correct representation of the world, which is not because it is just good enough for its purposes (Section 3.2).

phylogenetically inherited. They determine a life-compatible domain for physiological (e.g. heat, light, water, blood sugar levels), social (e.g. friendship, love, acceptance, trust), and general any variable (self-efficacy, competence, etc.) that is expedient to further one's own existence. In summary, they define what the creature believes to be and wishes and strives to become. In other words, "systems must behave in a way that increases the evidence for their own existence" [222, 104] or "you have to expect things of yourself before you can do them" (Michael Jordan).

The existence of top priors implies that only modeling the world, *even including the body*, is not enough. If our lives were ephemeral enough that we never witnessed the effect of our deeds on the world, as a particle submerged in a thermal bath, then we could get by with only a rough model of the environment. But instead, we are non-equilibrium steady state systems or creatures defined by our persistence through time (i.e. stochastic attractors [51], Sections 3.2, 5.5), and our existence requires a model of the environment to counteract noxious fluctuations that may arise as a consequence of the interactions between our actions and the environment, at all times [94]. Our actions in the present affect the world in the future, and our sensations in the present affect ourselves in the future: the brain and the environment are bidirectionally coupled, so one must model either both or none. So action and perception require the d-self aspect of self-knowledge, i.e. what and who you believe to be (Section 4.1.2). This entails that an agent can only entertain a self-consistent belief about its future behavior, i.e. that it will minimize (the time integral) of variational free energy [103]. Finally and importantly, although you need a good enough model of your own top priors, to predict how you will interact with your surroundings, both the generative model and its light diachronic version are just approximate representations, so one must bear in mind that the representation of top priors comprised in the d-self must also be a (good enough) approximation. This could explain e.g. the existence of hot-cold empathy gaps [182].

4.6.2 Uncertainty sources within the brain: forecasting one's own actions and thoughts

However, in general this is not true for active inference; specifically, in forecasting via diachronic modeling (Section 3.3). This is simply because diachronic modeling draws on a random generator to accomplish approximate optimization, and a stochastic source injects irreducible uncertainty into the system trajectory. Hence in some cases *not* building a model of one's own unpredictable forecasts or imagination could actually be more expedient or decrease more variational free energy than doing it. While the (for the most part unconscious) brain dynamics coasts on variational free energy, the uncertainty or entropy generated by the stochastic generator is subsumed under the d-self's intrinsic uncertainty. In fact, most of the p-self cannot and is not explicitly modeled, because it is mostly constituted by either not expedient enough mechanisms or irreducible noise. Crucially, the combination of forecasting relying on stochastic sampling (to render inference tractable) and noise not being modelable entails that decision making is unconscious and thus that the brain, not knowing the ultimate reason, has to constantly make up (infer) reasons for its own (the d-self's) behavior (Section 4.3.1).

Brain noise entails that predicting our subsequent thoughts or actions becomes exponentially improbable in a timescale of seconds. But note that even if the brain acted as a deterministic device, optimizing action would still be non-trivial: one needs to predict what one will do next accounting for one's own model of oneself (i.e. while knowing that

one is acting while attempting to predict oneself), in an infinitely recursive loop of nested selves, which may not lead to a fixed point. However, entertaining a simpified model of the p-self (like the d-self) is likely to end this loop after a few steps. Intuitively, the reason is that the d-self is a lossy representation that "leaks" information every time it is used to substitute the p-self[61]. Thus there is no need for entertaining a sequence of recursive nested d-selves: a single d-self is good enough.

4.6.3 Sleep and dreaming: pruning the tree and training in a hot simulator at night

Suppose that for some reason, a creature could periodically afford to shut off its brain from all sensory input. What should it do? Sleep and dreaming seem to be a good answer.

Variational free energy can be conveniently decomposed into one summand term that depends on sensory input (accuracy) and another that does not (model complexity) [101, 136]. We can ignore the first term while sensory input is missing. The second term quantifies the information distance (Kullback-Leibler divergence) between the model evidence of the generative model (of the world's hidden causes or objects) and the model evidence of the ensemble or recognition density (that approximately inverts the generative model given sensory input; Section 3.2) at a given timepoint. Assuming that the recognition model inversion is sufficiently accurate[62], the generative model evidence can be viewed as the prior model evidence before the system undergoes a learning span (daytime) and the recognition model evidence as the posterior model evidence (at the end of the day). Then, nighttime "learning" or sleeping would correspond to adjusting the (especially non-sensory) generative model parameters to better fit the shape of the recognition model after diurnal learning[63] [136]. Moreover, their difference is directly related to the difference of complexity or number of parameters of the generative model with respect to the recognition model[64] [105], which in turn are encoded by synaptic efficacy, in particular by the (excess) number of synaptic connections [96]. Intuitively, a judicious simplification the generative model leads to better model evidence because the probability distribution does not sprawl wastefully across parameter space regions that do not account for the sensations or epigenetic priors [18]. Thus, synaptic pruning[65] (simplifying the generative model) during sleep [283] can decrease free energy and thus increase fitness [136].

Hobson and Friston have proposed that top-down predictions are ignored during sleep because sensory prediction error units have been rendered insensitive through aminergic gating [136]. This effectively would shut off the higher areas of the brain from the sensorium, thus unleashed from the influence of prediction errors (except oculomotor

[61] Consider for example that recursive thinking about oneself does not come about naturally: it is difficult to think of oneself thinking of oneself. The recursive selves' levels cannot be many. The mechanics of variational free energy descent remain mostly unmodeled, and thus unconscious.

[62] The recognition model is in general incapable of exactly inverting the generative model [134]. However, here we assume that the inversion is accurate enough for fitness purposes; otherwise, the generative model would decay into an overly simplified mean-field approximation model [106].

[63] In terms of the expectation-maximization (EM) algorithm [72], daytime and nighttime adjustments would be analogous to the E and M steps respectively [134, 95].

[64] The generative model is in general more complex than the recognition model, which often (in variational methods) is just a factorized version of the generative model [61, 101].

[65] In the sense of cutting particular superfluous or undesired twigs, branches, or boughs. Not to confuse with evenly trimming the envelope of twigs.

proprioception relayed by the pons [56, 136]). The system, although deprived from sensations, continues striving to minimize free energy. The higher areas, set loose, can pursue their fantasies [134, 265], amplified by neural noise [114] or environmental perturbations.

Besides pruning of synaptic connections, variational free energy descent is also carried out by descending along non-sensory parameters encompassing learned memories and structures about the world model [136]. We also suggest that just as during wakfulness, parameter adjustment occurs via diachronic sampling inference while sleeping (Section 3.3). This would be akin to training in a stochastic simulation of the world. The source of this stochasticity is likely to lie in brain stem sensorimotor circuits (activation-synthesis hypothesis, [137]): the pontine-geniculate-occipital (PGO) system conveys information about eye movements from the brain stem to the thalamus and visual cortex [136] in both waking and sleep and PGO waves elicited by saccades in rapid-eye movement (REM) sleep reflect not (geniculate) visual prediction errors but (pontine) proprioceptive prediction errors [136].

On a high-dimensional rugged variational free energy lanscape (such as that of our world), a coasting particle is almost certain to slide into one of the many ravines or gullies (local minima) and get mired down, at least for some time (Section 3.3). And if the odds of finding a lower local minimum are high enough, the resources spent on plying stochastic diachronic inference will be warranted even in the absence of sensory input. We surmise that such inference is likely to comprise both perceptual *and* active components where the generative model plays out simulated stories and performs variational free energy minimization while the sensory input is effectively ignored (through aminergic gating that desensitizes sensory error signals [136]). So finding oneself in a night simulator with a atypical randomness (precision) configurations could be conducive to finding new solutions to problems that are more challenging with precisions optimized for wakeful inference.

During the REM stage of sleep brain activity is comparable to that of during wakefulness. This is consistent with REM sleep enacting simulated mutisensory and submotor world trajectories [142] where the inference engine can scour for new local minima (e.g. strategies to escape threatening situations and nightmares [223]), which typically requires the participation of the d-self. Conversely, during non-REM sleep metabolic energy consumption is reduced all over the body. Slow-wave sleep (SWS), the deepest phase of non-REM sleep, has been associated with facilitation of long-term memory consolidation via repeated reactivations of newly encoded information in the hippocampus, which mediates the transfer and integration of declarative memory with pre-existing knowledge on the cortex [77]. This is consistent with the hippocampus sitting near the top of the cortical hierarchy [196] and featuring adult neurogenesis [37] and with SWS being involved in complexity reduction through synaptic pruning: recoding unprocessed new declarative memories initially dwelling in a higher level (e.g. hippocampus) into lower level (e.g. inferior temporal cortex) concepts [128, 77] can be a way to cut down redundant connections [52].

Why do we seem to (or can recall) have dreamed only some nights? Typically, only dreams that are vivid and occur during or immediately before waking are remembered [137]. We propose that conscious recollection of dreams can only be associated with sleep periods that leave a memory trace from which a d-self can be reassembled *as if* it had been present in the dream. Of course, this is more likely to occur if the d-self had been actually invoked during the sleep. Dreams are mainly reported during REM sleep —although not only and not always [265]. The indispensable role of REM

sleep is consistent with non-REM sleep decreasing progressively since birth as waking time increases and REM sleep plateauing after its decline in the first year at about 1.5h for the rest of life [139]. Hence it is plausible that the brain carries out inference involving the d-self (or at least inference that boosts posterior dream recall) in a secluded regime of shut down sensory input chiefly during REM sleep. The setting of precision hyperparameters at different levels of the cortical hierarchy characteristic of different sleep stages (via joint aminergic and cholinergic modulation) [136] could explain their phenomenological differences. REM sleep features stronger emotions, remote or bizarre associations, conspicuous first person agency and motion in fictive space, but weaker self-awareness, metacognition, logical thinking, orientation, and memory than wakeful perception [138]; non-REM dreams tend to be more veridical then REM dreams [138].

4.6.4 Metacognition and introspection: reassembling memory pieces

Metacognition can be defined as "cognition about cognition" [87] or knowing what you know, which typically entails introspection, propositional beliefs, or scanning one's own experiences in a self-aware manner [213]. Despite our beliefs and its etymology, introspection seems not to have accurate access to decision making processes [207]. Ostensibly, we rummage in our minds for unconscious objects to bring them out to consciousness, as in remembering some hidden piece of memory. But memories are not, like pictures, factual portrayals of past events. Instead, they strike one as biased, fallible and untrustworthy *when* judged by its performance as a reliable store of information [235]. This is simply because expediency (free energy minimization) and accurate storage are different things (Section 3.6). Hence perceptual performance is typically, but not always [125], strongly correlated with metacognitive performance, which is usually measured via introspective judgments such as perceptual awareness or confidence [120, 251, 215].

However, it is possible to experimentally dissociate behavioral performance from confidence judgments [165, 244] and subjective experience [145] in normal subjects. The precision of neural representations is likely to be encoded by thalamic nuclei and its modulating effect conveyed by thalamocortical connections [256, 153]. However, the precision relevant to the d-self is (phenomenally) manifested as confidence [199], which is typically associated with ventromedial prefrontal and prefronto-parietal areas [86, 225]. Introspection is typically inaccurate, but that is normal: from an evolutionary perspective, what is sensible is not being aware of as much information as possible, but providing a faithful enough representation of the state of the world to assist survival.

We view metacognitive thinking as reassembling pieces of information pertaining to the d-self, that is as inference about the (retrospective, prospective, or fictitious) knowledge or actions of the d-self. This interpretation furnishes a simple and unified account of the mechanisms underlying subjective reports, that could adjudicate the current disarray of consciousness theories (Section 6). For example, consider the phrase "I think of what I did": we tend to identify the second "I" as the present immutable p-self, whereas actually all we can consider is an estimated reconstructed past d-self (a past sesmet, see Section 4.5.3). The past is not retrieved, but inferred. Metacognition is not simply retrieving static objects, but actively guessing what "I" saw, did, and learned from the past and currently unfolding piecemeal evidence. Hence, asking whether conscious percepts is dichotomous [246] or gradual [212] turns into asking whether percepts are ascribed (inferred to belong) to the d-self is an all-or-none manner. The answer is governed by a stochastic nonlinear process specific to each decisional context (such as the choice set

[192]) that can in general exhibit many diverse transitions between on and off states (where each component or feature of the percept can be independently be ascribed or not to the d-self [166, 44]) and thus requires description in terms of ad hoc probability distributions.

A common theme of discord is whether subjects can consciously perceive objects that they did not report, but which they *could have* reported. For instance, this is plainly illustrated by Sperling's paradigm [269] and its analogues in other modalities [25, 55], which are compatible with iconic or in general short-term memory being amenable to posterior conscious rekindling as a function of elapsed time [270, 307] and contextual cues. A subject that briefly "saw" all characters on an array, was cued and reported the characters on the bottom, during which the rest of characters were "forgotten" [269]. Here the impression of perceptual richness is afforded by the ability to reconstitute past or counterfactual events [166]. The moot point is whether subjects were ever conscious of the non-reported characters. Clearly the answer depends on what is meant by consciousness. For example, the reported characters are said [26, 166] to belong to access consciousness, and the non-reported but displayed characters to phenomenal consciousness (Table 1, Section 4). Consciousness is a suggestive yet ambiguous term, so there is a case for breaking it down into simpler components. However this can be detrimental if thereby we do not acquire some immediate clarification, because —as in the cups and balls or shell game— the pea that before lied under a known cup has now been quickly shuffled under one of two shells (Table 1). In fact, the common sense psychology meaning of consciousness is subjective experiences that are currently acknolwledged to pertain to one's own experience, and can be reported as such. These are precisely the bound phenomenal states (Table 1), which ensue from neural representations that are ascribed to the d-self (Section 4.3).

Objects are ascribed to the d-self depending on whether subjects believe that they saw a particular object, which in turn is the result of an inferential process or *guess* made by the brain based on piecemeal evidence [166]. Thus for example our inflated sense of perceptual richness and its attending overconfidence [209] may simply reflect that having a self-consistent d-self across time that believes it knows more than it does has been advantageous for fitness. Crucially, the inference is about whether the d-self, and *not* the p-self (which cannot be directly known), should be attributed some knowledge. This implies that even if an autopoietic system has evidence of the presence of a particular object in the surroundings, it may choose to ignore it if it is not relevant to its simplified model of itself (d-self) as having perceived it. If this is surprising, recall that the goal of organic inference is not accurately representing the world *per se*, but something that can be expressed as roughly equivalently expediency, free energy minimization, enhancing phenotype fitness, gathering evidence for states that are compatible with one's existence, or striving to survive. For humans, situations where accurate representation and expediency are at odds are typically resolved by reacting in an unconscious manner to sensory error signals. In brief, (you or the d-self) knowing what you have done and what you know is often useful and even indispensable, but it is by no means the purpose of biological function. You (the d-self) are invoked as the need arises, and likewise are granted knowledge and confidence about the world only when it is expedient. In this view, it may be debatable whether "the unexamined life is not worth living" [216], but it is plain that as long as there is no activated d-self to ascribe it to, life is not lived by any observer.

4.6.5 Mindfulness as a self-reverting random walk

Mindfulness is a meditation technique derived from Buddhist and Hindu traditions that can be defined as bringing one's attention to the present moment, on purpose and non-judgmentally or simply as "being here now" [151]. After stopping to contemplate the flow of experience without either acting on or thinking of what is the next step to take (forgoing active inference), one may gently hold in check the diffusing chain of thought by bringing it back to an anchor (e.g. breathing) or alternatively let go and observer non-judgmentally how imagination unfolds [16] (by coasting on the variational free energy landscape). It involves sustained[66] attention directed to oneself (d-self), but otherwise without disturbing the natural course of the chain of thought[67] (or spontaneous variational free energy descent). This can be visualized as a drifting mental random walk attracted by or reverting to states where the d-self is activated[68] —by deliberately keeping the d-self activated, and reactivating it when it has faded away.

Interestingly, awareness is encouraged to be brought to current experience and to where it wanders away in a particular manner: infused with mellow curiosity [20]. The reference to curiosity is not perfunctory: without a curious disposition, upholding an activated d-self would be virtually unfeasible. This is because the d-self is only needed so far as a problem is recognized whose solving requires its involvement (Section 4.5.1). Sensory and lower level prediction errors are either suppressed by readjustment of lower level parameters or are relayed upward. But the prediction errors (induced by online action selection) at the highest level that implicates active inference (corresponding roughly to the theta frequency scale of the brain; see Sections 6.6 and 7 of [191]) —which virtually always involve the d-self— cannot be relayed upwards. We speculate that this is because active inference is in general intractable, so there are no supraordinate levels to the d-self that can efficiently infer optimal action. Instead, diachronic inference sits at the highest level of action selection, at the level of the d-self (Section 3.3), to facilitate active inference via stochastic sampling, thereby generating actions that are attributed to our (the d-self) free will (Section 4.3.1).

The practice of mindfulness can reduce stress and anxiety and boost psychological well-being [24, 155] by encouraging a discipline of keeping a prudent attitude that strikes a balance between avoidance and excessive engagement of emotion or rumination. To wit, suspending judgment and intentionality can temporarily free us from the preccupation and anxiety [16] that goad us to minimize variational free energy without telling us how. In Buddhism, mindfulness is encouraged as one of the practices required to bring

[66] The term mindfulness derives from the Pali word *sati*, which means remembrance or memory, in the sense that one should remember to be aware or maintain the disposition of observing one's own mind and return to oneself (reactivate the d-self in our jargon) after having wandered off.

[67] Note the parallel between Husserl's phenomenology [292] and mindfulness: "bracketing" our assumptions about the world (i.e. holding back priors) to remove the idionsyncratic intentionality that constitutes objects in consciousness can be transcribed as observing non-judgmentally. It is also notable that people in the autism spectrum disorder may be more predisposed to mindful states owing to their "hypersensitivity and the reduced influence of cognitive priors" [204].

[68] In dynamical systems jargon, the mental random walk is attracted only along the dimensions of the (stable) subspace or manifold of states with activated d-self. In the absence of other (variational free energy) forces, the rest of dimensions would constitute a slow manifold where the random walk diffuses isotropically.

suffering to an end[69]. Daoism[70] emphasizes contemplation (mindful stillness) and open-mindedness (emptying onself, i.e. weakening priors) to achieve effortless action. This can be viewed as avoiding over-reliance on mechanical rule-based performance and stiff scheduling and favoring intution and "unconscious action". It can also be useful e.g. in dialoguing, to disregard the petty word nuances and arbitrary semantic dichotomies (in the structural linguistics sense) employed by sophists to equivocate and instead to focus non-judgmentally on the meaning of what one intends to get across [16]. Our thoughts emanate from expediency seeking or variational free energy minimization, but this source of "vital force" has been variously denoted positively as Dao or the Way [16], God, the Will, Dasein, etc. (Section 4.3.1) or negatively as anxiety of death [12]. For us, most of these realizations involve acknolwledging that our thinking is unconsciously generated and retrospectively attributed to ourselves (to our d-selves actually; Section 4.3.1) so assuming that the d-self is the source of ideas can readily lead to wrong conclusions. In other words, mindfulness can be seen as intuitively realizing that paradoxically my subjective experience can partially be detached from the d-self.

4.7. What I (d-self) experience is not the same as what the body (p-self) experiences

Let us conclude this long chapter with an elucidation of what introspection and subjective reports (mentioned in Section 4.6.4) are. We propose that —besides furnishing a sense of "being" (Section 4.2)— grasping that we mistake the d-self for the p-self can resolve many dilemmas that turn up in contemplating the determinants of subjective experience. Briefly, this is because the knowledge encompassed by the p-self and the knowledge attributed to the d-self are different things: the former is all information gathered by the living creature or autopoietic system, whereas the latter is just the experiences that are expedient to ascribe to the d-self in pursuance of optimizing behavior.

We suggest that (current) conscious experience is sensibly defined as the plain answer to the question: "What am I experiencing now?" A self-reflective prompt is essential because it induces the brain to invoke its representation of itself, i.e. it activates the d-self. This follows from the scheme offered in Section 4.3: the contents of subjective experience or consciousness, in the sense of bound phenomenal states, are defined by the d-self shell or set of deictic representations ascribed to the d-self at a given timepoint. This

[69] In Buddhism, humans being delusional about the three marks of existence is believed to result in suffering. The three marks of existence are fundamental characteristics of any being that humans typically are not aware of: impermanence (*anicca*), unsatisfaction, emptiness or non-self (*anatta*), and incompleteness or suffering (*dukkha*). In brief, any being is impermanent, incomplete or unsatisfied, and it lacks a permanent core that distinguishes it from the rest of beings. This inspired Schopenhauer to believe that the insight that individuality is a mere illusion is alleviates the anxiety stemming from the misconception of being a stable self [242]. Curiously, Pyrrho (Section 1) was aware of these ideas, and he rephrased and integrated them into philosophical skepticism [10]. The three marks of existence can be construed as an expression of the intermittent nature of sesmets or on-off d-selves (Section 4.5.3), our poor knowledge of the attributes and motivations that define and drive us (Sections 4.2 and 4.3.1), and the endogenous driving force (the gradient of variational free energy) whereby we ever feel some degree of pain or discomfort that impels us to take action, which seldom leads to a long-lasting satisfaction (which is virtually impossible in a complex and stochastic world; Section 3.2). Mindfulness enables to "keep up with the shiftiness of changing experience" [292], which is required to notice e.g. the fickle nature (anicca) of d-self intermittent activations or sesmets (Section 4.5.3), their imperfection as synecdoches or models of the p-self (Section 4.2), and our lack of insight into the origin of the driving states and emotions that govern our actions (Sections 4.3.1 and 3.3).

[70] Zhuangzi [306] (see Section 1) and Laozi [173] are the main references of Daoism.

is also related to the somewhat tautological but non-trivial observation by Kant that any conscious experience requires an observer. We argue that the many other possible definitions of consciousness are less sensible. For example, in common sense psychology, the term consciousness usually refers broadly to brain states that may straddle present, remembered past, and guessed past regardless of whether they coexisted with a d-self (i.e. belonged to sesmets) and were explicitly imputed to one's own (d-self) experiences. This stance risks of regarding a past (unconscious or not ascribed to the d-self) brain state that can be retrospectively reconstructed and experienced *now* as if it had been experienced *then*. Reiterated retrospective reconstruction is most of what perception is about (Section 3.2); this is akin to intermittent short-term amnesia or to a computer that loads into main memory the content of its previous session upon reboot, but with fair amounts of added speculation. Consciousness puzzles start popping up when retrospective inference makes the simplifying assumption that past unconscious (d-self absent) brain states were conscious (d-self present). The reason is —we speculate— that this is expedient by because it affords a continuous-time d-self construct that can be leveraged to more simply and efficiently account for the current state of affairs[71]. Importantly, this entails that invoking the d-self typically leads to a shrinkage of the complexity of the currently represented scene. At any rate, this sets the stage to consider past brain states that were never ascribed to a d-self or experienced consciously (but nonetheless were processed within the p-self) to be regarded as having been consciously perceived (within a past sesmet). For example, say a few minutes ago you were mindlessly scanning a crevice on a wall. Asking "What was I doing then?" compels the brain to infer what the d-self, and *not* the p-self, might have been doing then. When recalling what you were doing then, you tend to infer that you were doing something consciously, but in fact, *you* were not. In a sense, you (the d-self of the current sesmet) did not exist back then. By invoking the d-self now, the up to now unbound phenomenal states "become bound" or conscious (Section 4.3).

It seems unavoidable to bring up the "I" every time one ponders over subjective experiences[72]. This is a manifestation of the indissoluble coupling between subjective experiences and subjects, which depending on the viewpoint can be reduced to a tautology. Namely, if you assume that a given subject has conscious experiences, you will not find them anywhere else (i.e. we do not expect unbound phenomenal states or experiences lacking an observer to be meaningful; Section 4). A priori, phenomenal states need not be bound to a d-self, but in the practice of reflective thinking this is virtually unavoidable; as soon as "I" is involved (in the form of a deictic center), they immediately become bound[73]. The d-self is a represention of a concept just like that of any other such as a face. Disrupting the neural substrate of face representations simply leads to prosopagnosia or agnosia of faces. But disrupting the d-self at a specific timepoint implies removing the anchor of subjective experience (or bound phenomenal states) and thus prevents the existence of sesmets, along with consciousness as defined here. Any subsequent recollection involving a d-self of that (unconscious) moment must be necessarily reconstructed *post hoc*.

Casting inferred world objects in subjective terms has implications not only for con-

[71] For example, "I turned down the cake because I believe cow milk is unhealthy" or "because cake is bad" as opposed to "I don't know why I turned down the cake" or "Actually I didn't turn down the cake, but an inner unspeakable unconscious force did".

[72] E.g. "I think therefore I am" [75].

[73] Note this strongly resonates with higher-order theories of consciousness (Section 6.1).

sciousness but also for the information content itself of representations. This rests on subjective experience resulting from inferring what experiences should be attributed to the d-self in order to optimize behavior and perception (Section 4.3). The idea is that explicitly asking (enforcing an explicit representation) what my subjective experiences are typically leads to a narrowing the scope of those same experiences. Conversely, emphasizing less the focus on yourself as a subject (d-self), potentially can expand the scope of conscious experiences. This speaks to the important distinction between introspective or egocentric ("I see a magpie on the tree") and environmental or allocentric reports ("There is a magpie on the tree") [26]. In most ecological conditions we do not make spontaneous subjective reports about the world [26, 192] because we are not mindful (Section 4.6.5); under the current definition of consciousness the environmental reports can be unconscious (unbound phenomenal state), but the introspective reports[74] are always conscious (bound phenomenal state). In other words, the way in which the brain is cued to envision the d-self and its purpose in a particular context will shape (typically by shrinking) and thereby distort [172, 287] the "shell" of constructs ascribed to the d-self, which bear a one-one phenomenological mapping onto our subjective experience. The current definition also enables delineating a clean answer to the question of whether conscious objects are dichotomous or gradual phenomena [245, 166] (Section 4.6.4): they are events that occur with the probability that the object is ascribed to the d-self. It could also reconcile opposed views that conscious and unconscious processes are tightly correlated [78] or disjoint [264, 145]: percepts can be modified post-decision depending on how the cuing question was framed. Similarly, perceptual processing of "preconscious" [68] weak stimuli would remain unconscious and wholesome to the extent that subjects are not forced to make explicit inferences on them, thereby distorting and recasting them into a stiff mold of explicit categories [192]. Finally, we can construe the debate on whether the richness of perceptual experience overflows the richness of our knowledge or reports [49, 127, 210] as a question of how we elicit information from the creature, e.g. to which extent we emphasize introspective over environmental reports.

5. Physics and psychics (phenomenology), noumena and phenomena: two sides of one thing

We have so far described, from a given observer's viewpoint, what mechanisms could underpin the boundary that defines conscious and non-conscious objects. Let us now discuss the implications for philosophy of mind.

5.1. Self synecdoche: a too coarse model of the model itself to perceive and act on the mechanics of perception and action

Being baffled by consciousness is typically caused by contrasting the unfolding of subjective processes such as perceiving, pondering, and making decisions with the third-person perspective of conspecifics' homologous processes and intuitive (and book-learned) physics[75]. The gist of the conundrum is illustrated by the philosophical zombie (p-

[74] Note introspective reports are analogous to meta-representations, which are representations of representations in higher order theories (Section 6.1).

[75] For instance: "The problem of consciousness (or more correctly: of becoming conscious of oneself) meets us only when we begin to perceive in what measure we could dispense with it [...] For we could in fact think, feel, will, and recollect, we could likewise 'act' in every sense of the term, and nevertheless

zombie) thought experiment [160]. P-zombies are concocted creatures prescribed to behave exactly like humans but lacking conscious experience[76]. It seems *plausible* that p-zombies could exist. At least, the current state of affairs in physics (nor in any other field of knowledge) does not forbid it. But trying to ascertain under which conditions can p-zombies exist is similar to tackling the hard problem of consciousness [41] and easily leads to an impasse [190].

A more fruitful approach is perhaps first addressing the meta-problem of consciousness or asking why we[77] think that there is a problem of consciousness [43]. To think that there is such problem, one first needs to discriminate between first-person or introspective and third-person or environmental perspectives [26] (Section 4.7). In other words, to be able to contemplate a consciousness problem, a creature needs to entertain a model of itself or d-self because otherwise all its beliefs will be environmental (or "third-person", but without a first-person counterpart). In a system lacking a d-self or model of itself as embedded in the environment, it is impossible to formulate a contrast between first-person (subjective) and third-person (objective) states because it lacks the concept of perspective or grammatical person, only implicitly impersonal representations of world states. Notice that such representations would correspond to unbound phenomenal states within a "hollow observer" [190]. We mentioned in Section 4 that this resembles a paradoxical status (at least on the phenomenological side) where of objects are represented without the participation of any subjective perspective. However, our breakdown of living systems suggests that there is no contradiction. Contrarily, it hints that the only reason unbound phenomenal states seem paradoxical is that empirically there is a perfect correlation between first-person perspective and phenomenal states. This correlation agrees with our thesis that bound phenomenal states (the common sense meaning of conscious experience) appear when a representation is attributed to (some aspect of) the d-self at a specific timepoint (Section 4.3). In other words, things seem to become conscious not because their representations are activated, but because I —the observer or more precisely d-self— am activated.

We suggest that the root of the meta-problem is the differences between how we understand or model the world and how we experience or access it. Laying bare the illusion of reality (Section 1) reinforces this suggestion by providing an important[78] clue: we have never directly "experienced" the environment, in opposition to what common sense dictated before Zhuangzi, Carneades, Kant, etc. suggested otherwise (Section 1). The meta-problem stems fundamentally from an *epistemic rift*: first-person and third-person views are fully dissociated, so we have never seen a "thing in itself" or noumenon. Third- and first-person perspective are qualitatively different: the former is (indirectly) perceived, inferred, or believed but the latter is (directly) experienced. In particular, our

nothing of it all would require to 'come into consciousness' (as one says metaphorically). The whole of life would be possible without its seeing itself as it were in a mirror: as in fact even at present the far greater part of our life still goes on without this mirroring, —and even our thinking, feeling, volitional life as well, however painful this statement may sound to an older philosopher" [206].

[76] David Chalmers [42] exploited p-zombies to refute physicalism. Roughly, his argument says that given that the scenario of a world inhabited only by p-zombies seems plausible, it would follow that physics alone cannot account for phenomenal states (consciousness).

[77] This does not include the illusionists, who believe that there is no problem of consciousness or deny that there exists a subjective aspect to our perceptions [73, 276, 220] (see Section 6.5). Yet the disagreement between illusionists and realists over consciousness might be more semantic than fundamental.

[78] But not indispensable, because even if you believe that you have direct access to environmental objects, you still know that your thoughts and emotions are subjective.

intellectually learned physics third-person understanding of our selves (which we could call "b-self" from book-self) is inconsistent with our intuitive first-person understanding of our selves (d-self).

Hence, the pivotal point of this epistemic rift is that our generative model of the world has evolved to represent objects that are relevant to biological fitness, but these do *not* include the properties of its physical substrate —namely the brain, which is the only place where our experiences occur— that are relevant to neural computatation (Section 4.6.1). As far as the d-self encased in the skull is concerned, we only need to represent things occurring outside the skull (the only exception is perhaps stochastic diachronic sampling, see Section 4.6.1). This does not mean that we cannot perceive our own brains —we could work out the details, involving mirrors and a craniotomy, if we *really* wanted to— but that our sensory sheets and motor apparatus are not suited to infer the sort of properties (electrochemical signals at micro- and mesoscopic scale) that characterize neural computations and their attending phenomenal states. Although the computations carried out within the brain are indispensable for perceptual and active inference and their metabolic cost have a marked impact on homeostasis, the computations themselves have no bearing on the creature's fitness (Section 4.6.1). Hence, the electrochemical spatiotemporal dynamics of the brain, which is the infrastructure of the generative model, is mostly opaque to the generative model itself.

But how would a generative model with an inferential engine that had access and performed inference on all the physical properties relevant to neural computation of other brain or even itself look like? If we assume that inferring these computations is expedient for the observer system similarly to how it is for the origin system (which implies both have the same generative model), then both systems would end up displaying near-identical activity patterns (at least, if the observer manages to ascribe its percepts to its d-self)[79]. Otherwise, the observer system would try to make sense of the origin system's activity with its own, different, generative model, and extract different hidden causes, which differ from the origin system to the extent that their generative models differ. For instance, we are endowed with eyes but as autopoietic creatures we have little benefit from understading the electrochemical properties of the brain (or of most things, for that matter), so we can only perceive a dull and squishy lump of jelly. Conversely, if we could augment our sensory apparatus with e.g. a detector of magnetic fields, our qualia space would be correspondingly expanded, but it would still remain ineffable.

This thought experiment suggests that in principle, it is possible to perceive or feel and interact with the phenomenal properties of matter (brain), but for that we would need to interfere with brain activity in a more precisely targeted manner (at least at the mesoscopic neuronal circuit level) than with the current causal methods of intervention, mostly restricted to occasional intracranial recordings or stimulation during epilepsy surgery [190].

Intuitive physics can afford a knowledge of the macroscopic appearance, consistency, and texture of the brain, whereas book-learned physics affords a some description of the brain electrochemical signaling, in mathematical language. The first is a perceptual reconstruction of the appearance of the brain via reflected (on neural tissue) and absorbed and transduced (by retinal cells) light into electrical signals, whereas the second is a mathematical abstraction gleaned through painstaking application of the scientific

[79]Similarly, if two d-selves coexisted within the same generative model, their subjective experiences would not be private anymore.

method. But crucially, none of them affords a way to interact with the substrate of phenomenal experiences in a precisely targeted, and controlled manner [190].

The segregation of where our model and understanding the world abides (mesoscopic neural circuits within a bone chamber) and where and what in evolutionary terms is the target of our modeling or understanding of the world (macroscopic surroundings of the body) entails that by "design" we are incapable of intuitively or phenomenologically understanding brain operation. Conversely, despite book-learned physics affording some understanding of neurophysiology, it has no bearing on how this knowledge translates into phenomenal states. Appreciating that physics has no bearing on subjective experiences is typically baffling. This is most glaring for properties of subjective experience such as color or taste (qualia [80]) that bear no obvious correspondence to the physical properties they stand for. The prodigious achievements of physics in reducing much of the world we know to a set of mathematical rules has so far been inconsequential for the puzzle of consciousness.

In summary, our current book-learned and intuitive knowledge of physics is consistent with phenomenal states being contingent epiphenomena suffusing physical matter, but also with physical and phenomenal states being two inseparable aspects of the same thing. This is in essence because our generative model of the world is unfit for (but not necessarily incapable of) understanding the relevant physical properties of its own substrate (and of our d-selves), which are associated with its attending subjective experiences.

5.2. Ontological uniformity: panpsychism as physicalism via realistic monism

Within the array of camps in philosophy of mind, the discussion so far has aligned us with philosophical skepticism [190] (true belief is impossible, but reasonably accurate belief is expedient) but also with realism (the thesis that objects exist autonomously or independently from observers).

Although some people deny the existence of consciousness or regard it as something already intuitively understood or explained by our current knowledge [276], it seems farfetched to believe so[81] except perhaps by dint of a semantic gimmick. In fact, the mismatch between our generative model target and its physical substrate or brain (Section 5.1) implies that arguments from many of the philosophy of mind camps over consciousness are at high risk of degenerating into semantic disputes akin to whether the two sides of a coin are two different things, two aspects of one thing, or actually just one thing that seems to be two different things. Currently physicalism is (transitorily) an incomplete thesis as long as physics does not account for the mapping between physical states and subjective experiences.

[80] In the Frank Jackon's [146] sense of "features of the bodily sensations especially, but also of certain perceptual experiences, which no amount of purely physical information includes".

[81] "That a world-interpretation is alone right by which you maintain your position, by which investigation and work can go on scientifically in your sense (you really mean mechanically?), an interpretation which acknowledges numbering, calculating, weighing, seeing and handling, and nothing more—such an idea is a piece of grossness and naïvety [...] Would the reverse not be quite probable, that the most superficial and external characters of existence —its most apparent quality, its outside, its embodiment— should let themselves be apprehended first? [...] I say this in confidence to my friends the Mechanicians, who to-day like to hobnob with philosophers, and absolutely believe that mechanics is the teaching of the first and last laws upon which, as upon a ground-floor, all existence must be built. But an essentially mechanical world would be an essentially meaningless world!" [206].

We do not have enough evidence to sketch a detailed mapping between physical or noumenal and phenomenal states. Kant did not even attempt to explain what noumena are. The absence of clues and methods to broach this barrier has led some people to put forward bold or implausible theories to link noumena and consciousness with other enigmatic notions, such as conscious volition with physical forces [241], consciousness with quantum mechanics[82] [295, 299, 124], or conscious reasoning with non-computable algorithms [184] (but [59]).

But if we adhere to the common sense notion of consciousness (i.e. bound phenomenal states associated with representations ascribed to the d-self; Section 4.3), the current evidence is consistent with the physical-phenomenal mapping being such that neither every phenomenal state is mapped to by at least one physical state (since all unbound phenomenal states are inaccessible) nor every phenomenal state is mapped to by at most one physical state (because there seems to be an equivalence relation between systems performing the same computations, i.e. many ways to perform the same computation). Or in mathematical terms, the physical-phenomenal mapping is neither surjective nor injective. For example, the continuous jiggling of thermal (and quantum) fluctuations seems not to manifest in phenomenal states, but sensory and transcranial magnetic stimulation, and pharmacological interventions such as anesthesia have obvious effects on phenomenal states.

Functionalist views[83] of cognitive science advocate some sort of injective map from physical states to cognitive (and conscious) states, such as Chalmers' Principle of Organizational Invariance[84] and Maudlin's Supervenience Thesis[85]. Hence, if we restrict ourselves to bound phenomenal states (to avoid non-surjectivity) and aggregate any system configurations that perform the same information-theoretic computations or functions (to avoid non-injectivity), then there seems to be a one-one (bijective) mapping between physical and phenomenal states (i.e. between neurocomputational and bound phenomenal states). The main implication from this is that the information content of conscious states (or bound phenomenal states or attributes of the d-self) must be the same as the information content of their corresponding neural computations.

Given the evidence that both the physical and phenomenal are aspects of reality, and that they happen to be revealed through an epistemic rift where first-person and third-person views are fully dissociated (Section 5.1), the simplest explanation would be monism, i.e. that they are two ways in which the same thing manifests [190], and in particular realistic monism [275] (or structural monism [108]). Based on functionalist views, an *reductio ad absurdum* argument (although we contend the adjective) has been set forth that either computation is neither necessary nor sufficient for cognition or panpsychism[86] is true [19]. Its author interpreted it as a case against the former, but we assess it as en endorsement of panpsychism[87]. In the spirit of our distinction between

[82] Although the characteristic timescales of neural and quantum processes are widely off [280].

[83] Inspired by the development of the digital computer in the middle of the 20th century, Putnam[219] famously brought on machine-state functionalism, which is the thesis that cognitive states are functionally specified, and that cognitive processes are computational operations on cognitive states. But later he backed down.

[84] "Any two systems with the same fine-grained functional organization will have qualitatively identical experiences"[41].

[85] "Two physical systems engaged in precisely the same physical activity through a time will support precisely the same modes of consciousness (if any) through that time"[195].

[86] In this article we use panpsychism encompassingly to also mean panprotopsychism [116].

[87] Although we still agree with computations not being sufficient for mind. Setting out to spell out the

bound and unbound phenomenal states, monism is just the view that physicalism and panpsychism (the view that phenomenal or mind-like properties are an ubiquitous aspect of reality) are two points of view that reflect different aspects of the same reality. This is also called more precisely realistic monism [275]. Quoting Strawson: "But how can experiential phenomena be physical phenomena? Many take this claim to be profoundly problematic (this is the 'mind-body problem'). This is usually because they think they know a lot about the nature of the physical. They take the idea that the experiential is physical to be profoundly problematic given what we know about the nature of the physical. But they have already made a large and fatal mistake. This is because we have no good reason to think that we know anything about the physical that gives us any reason to find any problem in the idea that experiential phenomena are physical phenomena" [275].

5.3. A failed foray into the physico-phenomenal mapping

How can we concretely go about uncovering how physical states relate to subjective experiences? Qualia (in the sense of Jackson, Section 5.1) are typically phenomenal states that express an hidden combination of physical variables. For example, colors as spectrophotometric measures of retinal cones or hunger as the effects of ghrelin hormone release, triggered by low blood sugar levels, on the stomach. More generally emotions can be construed as interoceptive inference on visceral and somatic states [248]. In any case, the role of these representations is not to maximize physical information, but selecting the fraction that improves the host's fitness (if attributed to the d-self). Colors are a sensible research target because their input source (retinal cones) are fairly well characterized and amenable to measurement, but nonetheless they ineffable experiences.

Colorimetry, the science of color perception, describes which properties of the light incident on the eye result in different perceived colors, but not how to tweak neural circuits to cause an observer to see what we wish. We mentioned that the information in the phenonomenal properties of a normal or trichromat (having three color channels corresponding to types of retinal cones) human's perception must enable to discriminate colors in a similar way as an inference engine with access to the same three channels would (Section 5.2). If we could swap two of our three retinal cones or shift their wavelength response functions (cf. "inverted spectrum" thought experiment [181] and shifted qualia [36]), we would probably just see swapped or "shifted" colors. Presumably we would be taken aback in the beginning, but to the extent that our memories adapted to the new channel configuration, we would get accustomed and eventually even stop noticing. But colors are qualitatively different in an unspeakable manner, so this would be impossible to test other than via subjective reports.

The range of perceived colors afforded by human trichromatic vision corresponding to distributions of wavelengths in the electromagnetic visible spectrum can be illustrated by a two-dimensional domain or palette of chromaticities (cf. CIE 1931 color space) where combinations of two variables (roughly hue and saturation) yield one color (after factoring out brightness). This shows that we see a tiny fraction of all the colors that we could in theory see with more channels, which in principle could be an almost infinite-dimensional space. According to Section 5.2 it would follow that a tetramat (a person with four color

case for panpsychism (or the related neutral and realistic monism) is beyond the scope of this article, but the interested reader can refer to previous work [275, 190].

channels[88]) could actually experience more colors than us. More information content implies a larger set of phenomenal states (Section 5.2). But again, we cannot test this yet (we can test differences in discriminability, but not in experience) because of the epistemic rift (Section 5.1). However, let us suppose that we could overcome the epistemic rift by means of an ingenuous experiment in the future. Here is what we foresee would occur. We would be able to discriminate and *see* many more (ineffable) colors. But we would not be able to directly compare them to how we saw color in the past because those experiences belong to past sesmets —which are the minimal units of experience (Section 4.5.3)— in the same way that we cannot compare our experiences with the experiences of others. At heart, this is because comparisons between (physical or phenomenal) representations are possible only within an inferential system that encompasses both. In other words, I cannot see what you see because I am not you. The basic reason is that the d-self shell is the fundamental defining factor of the contents of (the common sense notion of) subjective experience (Section 4.3). In this view, panpsychism holds in the sense that non-conscious states are simply unbound phenomenal states. This leaves only a possibility for two subjects to experience the same objects: that both belong to the same inferential system. Although this seems theoretically possible, it is evolutionarily implausible.

5.4. The forces of thought as an expression of the forces of nature

We have mentioned in Section 4.6.5 that thoughts emanate from the endogenous forces that an autopoietic system should exhibit by the mere fact of its existence [101], and that the source of this forces has been referred to by different people as the Dao, God, or the Will. Given our instinctive urge to explain away as much as we can, it is unsurprising that many people have endeavored throughout history (and prehistory) to discover unifying causes for the hidden forces or causes inhabiting the world. Folklore, philosophy, religion, and physics evince different aspects of this endeavor, and have brought forth their respective answers. For example, mythological cosmology, Brahman in Hinduism and God in many religions are similar in spirit to a monistic identification between physicalism and panpsychism, and hint at the existence of unbound phenomenal states; noumena or the "things in themselves" look like matter in our intuitive understanding of physics; the Dao, the Will[89] and Dasein are analogous to the unconscious "forces of thought" derived from variational free energy gradient (Section 4.3.1); conscious states can be seen as the phenomenal aspect of the forces of thought restricted to d-selves (Section 4.2). In general, these various notions can be reconciled by realistic monism (Section 5.2) and the self-evidencing physical forces (derived from variational free energy gradient) that existing things must exhibit in nature [97].

[88] In fact, the common ancestor of vertebrates was a tetrachromat [148], and some humans are likely to be tetrachromats [150].

[89] Schopenhauer's metaphysical voluntarism, deems the Will as the essential reality behind the world as representation or "reconstructed perception". Each element of the multiplicity in which the world manifests has the same blind essence striving towards existence and life, and human rationality is not actually different from the rest of nature at the fundamental level [241].

5.5. On-off synecdoches as intermittent localized mirrors where the world reflects itself

The living creatures we know of are finite and localized in space and time, i.e. p-selves. This is evinced by their spatial (e.g. membrane, skin) and temporal (birth and death) borders, by the limits of the set of physical or phase space states —a stochastic attractor [51]— that characterizes them, and when represented as a Bayesian network by the set of nodes that render the internal or "own" nodes of the creature conditionally independent from the rest of the world (a Markov blanket [102]).

The membrane of the creature (p-self) contains a representation or synecdoche of its world and of the creature itself (d-self). What determines the boundaries of the d-self? That is, what sets the limits of the d-self shell? The centerpiece of the consciousness riddle is perhaps what sets apart conscious from nonconscious objects. In Section 4.3 we proposed that the boundaries of consciousness are defined by the attributes of the d-self (d-self shell) that maximize expediency (minimize variational free energy). In short, our proposal says that to explain the semblance of consciousness we do not need to invoke the concept of consciousness, but the model of oneself or d-self. This explains that despite being a superset of the d-self, the p-self as a whole is not conscious in the conventional sense (of bound phenomenal states, Section 4) because only a small fraction of it —the d-self shell— is arranged as a (egocentric) deictic center ("this is me") to which attributes can be anchored within a simplified model of the world and of itself. Roughly speaking, consciousness (in the sense of unbound phenomenal states) is always "out there" permeating the world[90]; instead, it is me who is absent most of the time, as an intermittent entity that is recalled or "remembers" and is retrospectively reconstructed each it comes into being. This perspective everts (Section 2) the typical approach to consciousness riddles by bringing focus on the self rather than on subjective experiences.

Hence, the concept of d-self shell (Section 4.3) can explain the consciousness boundary problem in the context of realistic monism, which harmonizes physicalism with panpsychism. Thus the notions of unbound phenomenal state and panpsychism are justified by the d-self shell identification with phenomenal states ("the very notion of self-consciousness presupposes that there is an alternative (non-self) consciousness" [104]). In other words, phenomenal states lie outside outside our sight (are unbound) because the brain "considers" that under its generative model, it is not expedient that we see them. This feature is motivated solely by what can be modeled and what minimizes variational free energy in the current context of perceptual and active inference (Sections 3.2, 4.3). Finally, it is ironic that the very act of asking oneself about one's own perceptions or memories can alter this self referential process by expanding or more typically contracting the d-self shell and its corresponding bound phenomenal states (Section 4.7). In other words, loosening the grip on oneself by e.g. suspending judgment in meditation can lead to a subjectively expanded range of experiences [16] (Section 4.6.5), which is consistent with the d-self substrate being disrupted in schizophrenia and by dissociative drugs (Section 4.3.4).

[90] Yet it is misleading to say e.g. "the brown onion is conscious or has consciousness" because it has no d-self. The onion is no more conscious than its husk or any part of it: the only reason that we choose to single out the onion from its environment is that it is a useful distinction *for us*.

6. Comparing theories of consciousness

Here we briefly review some of the many theories of consciousness [251, 259, 250] and compare them to the on-off or intermittent synecdoche theory. So it is now befitting to attempt a characterization of consciousness, in the sense of bound phenomenal states (Section 4.7), under the on-off synecdoche (OOS) scheme, before coming back to it in Section 6.6.

Given a living creature, its consciousness is (an expression of the information content of) the attributes of a simplified and intermittent representation or model (d-self) of the creature itself in the form of finite spatiotemporal domains of the world instantiated by the creature's generative model of the world. A few vital notes: (1) the living creature's behavior can be outwardly explained solely by endogenous forces striving to further its own existence, i.e. by trying to minimize its variational free energy; (2) the d-self has the attribute of believing (believes) to be the creature itself (p-self); and (3) this together with realistic monism entails that consciousness does not have any function but it is just one of the phenomenal aspects of the physics of autopoietic inference.

6.1. Higher order theories

Higher order theories (HOT) assert that mental states are conscious when they are the target of a meta-representation or higher-order representation that represent oneself as being in particular (first-order) mental states [175]. For example, in higher-order thought theory, the thought "I see a magpie on the tree" is a meta-representation or re-representation of the thought or representation "There is a magpie on the tree". OOS resonantes with parts of perceptual reality monitoring, where conscious perception arises when a higder network judges a firstorder representation to be a reliable reflection of the external world [174]; the self-organizing metarerpresentational account, where consciousness is the brain's (unconscious, embodied, enactive, nonconceptual) theory about itself [48] (cf. Section 4.6.1); and the higher-order state space theory, where introspective subjective reports are metacognitive decisions about a generative model of perceptual content [85].

HOT is currently the closest theory to on-off synecdoche; indeed representing oneself as having some experience sounds similar to ascribing a experience to the model of oneself or d-self or to shifting from allocentric to egocentric viewpoints (Section 4.7). However, this conceals a (perhaps) fundamental distinction: while HOT imputes consciousness to meta-representations, in OOS consciousness is *not* entailed by meta-representations[91], but rather is a particular configuration of the d-self, i.e. the set of representations that are currently being attributed to the d-self. We are aware that by twisting the semantics one could perhaps make an identity out of this distinction. However the following suggests otherwise.

First and most importantly, OOS provides an account of the observation that conscious experiences are always subjective or first-person, whereas HOT either inverts the causality arrow by prescribing that this observation —representing oneself as having an experience is the common factor to all subjective experiences— is what engenders consciousness, or simply states their association in a descriptive fashion. This follows from the broader definition of consciousness as unbound phenomenal states afforded by OOS,

[91] At least in the HOT sense that meta-representations are the causal factors that render their targets conscious.

there is no notion of "consciousness advent" or "ignition" [67]. Instead, a "window" is opened that looks out on a panpsychist world, where every state has an associated conscious side. Crucially, under OOS the scope of this window is specified by the attributes of the d-self (or d-self shell in Section 4.3) that are evolutionary expedient (or equivalently minimize variational free energy), thus providing a testable hypothesis.

Second, OOS emphasizes an essential distinction between the living system (p-self) that entertains a generative model of its world, and the subset of that model that represents the living system itself (d-self). This distinction is indispensable to enable the realization that we are ("I am") the d-self and not the p-self, without which one readily falls into the category mistake that our conscious beliefs and sensations are an exhaustive expression of the living system's internal states instead of simplified representations attributed to one of its subsets, the d-self. In contrast, HOT does not seem to distinguish between the d-self and p-self: most descriptions refer to "oneself" without further specification. This reveals a fundamental difference between the two theories. Most versions of HOT do not require that conscious perception be necessarily associated with a conscious metacognitive state; instead for meta-representations of second-order to be conscious (e.g. self-referential conscious thinking), they themselves must be the objects of some third-order representation [175, 250], and so on in a theoretically infinite iterative loop. Although OOS also allows for iterative thinking, it plays no role in the appearance of consciousness, instead being just an ordinary conscious object among all possibilities. On HOT's view, both "ones" seem to refer to the same concept, which coupled with the assumption of meta-representations causing consciousness could lead to infinite iterations. But on OOS's view, "one represents oneself in a particular way", only makes sense if the first "one" refers to the p-self and the second "one" (in oneself) to the d-self. Thus, e.g. representing an object as being perceived by oneself (i.e. attributing an object to the d-self, in our terms) never requires more than one step (Section 4.6.1). In other words, HOT is what consciousness would look like for a d-self that beliefs that it is the p-self.

Third, much of HOT's support stands on evidence implicating anterior cortical areas in conscious contents, especially the prefrontal cortex [175, 88], because these areas are associated with metacognitive judgments. However, it is plausible that anterior areas are simply involved in executive control and subjective report [287], which crucially requires retrospective reconstruction of past events. Conversely, under OOS consciousness is associated with the d-self shell, which is typically associated to deep medial brain regions ([40]; Section 4.1.2).

In summary, HOT is a descriptive theory mostly consistent with evidence in the same way that OOS is, but it is not derived from first principles so it lacks a mechanistic explanation[92] for the appearance of consciousness that OOS provides.

6.2. Global neuronal workspace, attention schema, and re-entry theories

The global neuronal workspace theory (GWT) proposes that mental states become conscious when they are broadcast or rendered globally accessible within an antomically widespread "global workspace" typically subserved by frontoparietal networks [65, 193], so it relies for validation on studies that associate consciousness with neuronal signatures of ignition and global correlations [193].

The two main features of GWT, ignition and global access, ensue from OOS's principles. Broadcast occurs via nonlinear "ignition", where recurrent processing amplifies and

[92] This is called phenomenological model in physics.

sustains a neuronal representation [70] (Table 1). Ignition is analogous to activation and upkeep of representations as long as they are expedient (first and fourth items of Section 3.6), which when coactivated with the d-self for some timespan (or sesmet, Section 4.5.3) lead to the common sense notion of consciousness (Section 4.3). In GWT, global access enables conscious states to adaptably select context-dependent behaviour. This is precisely one of the consequences of the OOS's tenet that the representations attributed to the d-self are those that further efficient diachronic inference at the macroscopic resolution of the d-self (Sections 3.3, 4.3). This entails e.g. that working memory information decays exponentially if not selected for ignition (or in OOS terms, as a d-self attribute); one fo the reasons this occurs is limited computational resources, which is evinced in the brain using a single hypothesis (variational mode; Section 4.4) per representation.

In attention schema theory (AST), the brain is able to selectively focus its processing resources (attention) and it holds a representation (attention schema) descriptive of attention [119]. This attention schema provides conscious access to objects. OOS and AST share one important feature: both appeal to the construction of *simplified and imperfect models*, respectively of diachronic inference in OOS and of attention in AST. These are closely related notions, because the simplifications that are expedient for performing diachronic inference with a coarse model of the self or d-self in OOS are also likely to be advantageous for attention control in AST, where "The construct of subjective awareness is the brain's efficient but imperfect model of its own attention" [119].

Re-entry theories associate conscious perception with top-down recurrent signaling [170]. In the local recurrency theory, Victor Lamme proposes that localized re-entrant processing within sensory regions typically can lead to consciousness [172], while frontoparietal regions are required for reporting subjective experiences [287]. This definition of consciousness is analogous to unbound phenomenal states and with the concept of preconscious states (see Table 1). This is related to how the appearance of (pre)conscious objects becoming conscious after the stimulus or sensation has ended [68, 247] can be explained as a transition between unbound and bound phenomenal states through retrospective inference or reconstruction (Section 4.7).

6.3. Counterfactual richness of thick temporal models and Markovian monism

Markovian monism asserts that all (autopoietic) systems possessing a Markov blanket have properties that are relevant for understanding the mind and consciousness [108]. It goes along with structural realism (which perhaps is the same as realistic monism [275], Section 5.2) or the view that the mathematical structure characteristic of the internal states of an (autopoietic) creature cloaked in a Markov blanket determines the relationship between (probabilistic) beliefs and the (statistical) physics of internal states that represent those beliefs, without making strong ontological commitments other than suggesting that there is a gradual difference between non-conscious and conscious entities [108].

In this scheme, agents that have evolved thick generative models [104] (cf. diachronic thickness, Section 3.3) are more plausibly labeled as conscious, because they have beliefs about what it is like to act [104]. According to this view, in non-conscious processes action selection is realized in the here and now, but conscious processes accomplish prospective inference, i.e. simulate multiple futures under different actions, and select the best action [104]. In other words, Friston argues that consciousness is brought about

by the counterfactual depth [254] accompanying thick temporal models[93].

In brief, Markovian monism with provides an account of much of the information structure characteristic of conscious experiences: being conscious is like being a persisting autopoietic system. However, Markovian monism attributes the difference between the conscious and non-conscious internal and blanket (particular) states [108] to the temporal thickness of the agent's generative model, whereas OOS attributes it to a selection process based on the expediency of the d-self and its ascribed representations (Section 4.3).

6.4. Integrated information theory

Integrated information theory (IIT) proposes that a version of a measure of the complexity of neural interactions that roughly quantifies the average mutual information between bipartition halves sum at different scales of a system [284, 107], called integrated information, measures the "quantity of consciousness" and the configuration of its corresponding irreducible (in a mutual information sense) "cause-effect structure" determines conscious contents [282].

IIT and OOS coincide in two points (which are common to other theories of consciousness): adherence to realistic monism and thus to panpsychism (only in the sense of unbound phenomenal states, Section 5.2; or structural monism, Section 6.3); and their emphasis on the unity of consciousness [154, 11] (denoted as unimodality or abductive inference in Section 3.6). However, under OOS (and Markovian monism, Section 6.3), the relatively high complexity of brain function is simply a reflection of the complexity of the world we inhabit [107, 46] (an expression of the self existence bias [190]), and has no other special relation to consciousness.

6.5. Illusionism and the meta-problem

Illusionism, most adamantly advocated by Daniel Dennett [74], is an eliminativist theory about (bound) phenomenal states: according to it, we do not actually have phenomenal states but merely represent ourselves as having such states [74, 220, 250], and even qualia do not exist [73].

At first sight, it seems that Dennett denies consciousness and nothing could be further from realistic monism and even common sense [276]. But on reflection, perhaps illusionism and realism are just using the word consciousness for different meanings. Illusionism holds that once it is explained why people believe and say they are conscious (the meta-problem of consciousness, Section 5.1), the hard problem of consciousness (Section 1) will have been dissolved [74] because "we are robots made of robots [...] who manage, in concert, to create a user illusion of a Conscious Person, a single, unified agent, a self" [74]. In fact this sentence can be rendered as an expression of OOS by only swapping "robots made of robots" with "p-self" and "illusion" with "d-self".

So perhaps what illusionism refers to as "illusion", which denotes non-existence, in some sense is actually meant to refer to the same thing that OOS refers to as "model", which is a very real both physical and phenomenal entity, just not the same as a p-self

[93] The idea is that counterfactual hypotheses would be different from the short-sighted perceptual hypotheses: a hierarchical generative model leads to the loss of phenomenal transparency (i.e. appearance of consciousness, or bound phenomenal states in this article) [197, 104] because the beliefs (probability densities) are not propagated or shared among the levels and their statistics (variational modes, Sections 3.2, 4.4) are available only to higher levels. This means there is an opportunity for "mental action" that does not entail overt action, which would render beliefs phenomenally opaque (conscious) [197, 104].

(Section 4). Thus, from OOS's stance illusionism would only apply in the sense of the user illusion, the notion that consciousness appears to be a simplified version of reality akin to a computer desktop [208]. Under OOS we can answer the question "why do people believe and say they are conscious?" by saying that the confusion comes from identifying "people" and "they" whereas in fact we should be careful to specify that they are different things: p-selves contain models of themselves or d-selves that believe they are p-selves, which in truth they are not. But of course we cannot avoid believing so because that is precisely how we (d-selves) have been defined (Section 4.2).

6.6. On-off synecdoche theory

Here we confront OOS with some questions any comprehensive theory of consciousness should provide answers to [250].

Why are some organisms or systems conscious whereas others are not? This is a loaded question. The answer depends on the definition of consciousness. If consciousness refers to unbound phenomenal states (Section 4), then all systems are conscious. But under the common sense meaning of bound phenomenal states (Section 4.7), the answer is that for a system to be associated with consciousness, it must possess a (generative) model of the world complex enough to comprise a representation of itself (d-self) to whom perceptions or knowledge can be attributed.

Why do states of consciousness differ from each other in their content (or local state) and level (or global state)? The information content of conscious objects is equivalent to the information content of the object's physical substrate —neural computations subserving the attributes of the d-self— which self-organizes into a unstable poise between competing top-down priors and bottom-up evidence [98] that can be disrupted pharmacologically (e.g. psychedelics). Consciousness levels are associated with subjective profile, arousal and behavioural responsiveness [250]. Under OOS, they are associated not with the specific attributes of the d-self, but with the activation of the d-self itself and its degree of relevance (expediency) for the current task set. We speculate that this mechanism is disrupted in schizophrenia and by dissociative drugs (Section 4.3.4).

Does consciousness have a function? No, but its existence is a manifestation of the utility or expediency of entertaining a model of oneself (d-self) and its content is a manifestation of which of its (d-self) attributes it is expedient to invest resources in modeling and computing (Section 4.3).

7. Conclusion

On-off synecdoche is a deflationary theory of consciousness: it regards subjective (self-centered and anthropocentric) experiences as being subjective not by virtue of "our own discernment" but of being "exogenously" (out of "me" or the d-self, but within the p-self) specified to be a simplified model of the host system or p-self (Section 4.2). In particular, this simplified model or d-self is incorporated in the architecture and dynamics of the host brain, which in turn are governed by a "vital force" or variational free energy gradient (Section 4.3).

On-off synecdoche hinges on the observer bias, the observation that having phenomenal experiences and asking questions about them is always conditioned on the existence of an entity who somehow[94] "owns" phenomenal states [190]. This observation is explained

[94] "We are somehow immediately presented with our experiences" [43].

by identifying this entity with the d-self or model, *not* the p-self or creature (Section 4), where the d-self is specified as just a brain representation with no more causal or free will powers than any other (e.g. Sections 4.3 and 4.7). Thereby our entrenched (default) belief that consciousness is a private property is dismissed in favor of consciousness being an omnipresent property of the world (unbound phenomenal states) of which we can experience a small fraction (bound phenomenal states) defined by the attributes, selected purely on account of expediency, of a simplified model of ourselves (d-self).

Standing on basic principles entailed by the existence of living systems and on phenomenological evidence, we inferred necessary properties that must feature subjective experience (simplicity and expediency, estimability, unimodality or abductive inferece, memory intermittency and decay; Section 3.6) and also sufficient conditions for the realization of subjective experiences (d-self specification, demarcation of the d-self shell by expediency or variational free energy descent; Section 4), i.e. for "binding" phenomenal states. These features are naturally incorporated into living, autopoietic, systems as a generative model (or synecdoche of world and itself; Section 3), an ostensible "vital force" governing behavior (variational free energy landscape descent; Section 3.2), a diachronic (stochastic, simplified and tractable) inference engine that accounts for the semblance of free will (Sections 3.3, 4.3.1), and a d-self within the generative model that represents a simplified model of the system itself (synecdoche for p-self) including the d-self itself (Sections 4.6.1, 5.1), which is intermittently turned on and off as the need (expediency) arises (on-off sesmets, Section 4.5.3).

On-off synecdoche is consistent with behavioral, clinical, and phenomenological evidence (Sections 3, 4, 5) and with the free energy principle (Sections 3.2, 3.3), provides a general framework with which different theories of consciousness can be collated (Section 6) and makes two key conceptual distinctions (1) between the p-self and d-self (Sections 2, 4) and (2) between unbound and bound phenomenal experiences (Sections 4.3, 5). It also yields a testable prediction relevant to subjective experiences: the d-self is just a neural representation, whose substrate is likely to be typically distributed along the posterior medial brain (Section 4.1.2), in particular in the vicinity of the precuneus [40], and the common sense notion of consciousness corresponds to the d-self shell, which is anchored in the d-self (Section 4.3), so e.g. mechanical or electrochemical disruption or deactivation of the d-self would lead to unconsciousness. However, it is crucial to note that the sort of "conscious entity" being disrupted corresponds to spatiotemporally localized and finite representations (on-off sesmets, Section 4.5) that are highly volatile and adaptive, in the sense that they are continuously being activated, retrospectively reconstructed, and deactivated. This means that it is intrinsic of subjective experiences, when depicted with respect to time, to be intermittent with a characterstic period of a subsecond scale (Section 4.5.3), so experimental interventions would need to last longer to yield appreciable effects.

The chief precept in conceiving on-off synecdoche was to balance simplicity and consistency with the current physiological and phenomenological evidence. However, it cannot answer the hard problem nor its related puzzles (Section 5.3), such as how physical states map into phenomenal states. These questions cannot be adjudicated with armchair arguments because first we have evolved to just well enough solve problems relevant to furthering our existence (Section 3) so any question we happen to be capable of tackling beyond that is a bonus; and second it could be a contingent property of our world[95].

[95] Perhaps like the fine-structure constant in physics.

Answering such questions will require at least hard evidence from systematic, controlled and neural-circuit scale precision brain stimulation experiments that currently are mostly unavailable (Section 5.1).

8. Acknowledgments

This article was supported by the HSE University Basic Research Program.

References

[1] Rick A. Adams, Stewart Shipp, and Karl J. Friston. Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218(3):611–643, may 2013.

[2] Rick A. Adams, Klaas Enno Stephan, Harriet R. Brown, Christopher D. Frith, and Karl J. Friston. The Computational Anatomy of Psychosis. *Frontiers in Psychiatry*, 4:47, 2013.

[3] John R Anderson. *Cognitive Psychology and its Implications, Sixth Edition*. Worth Publishers, 6th edition, oct 2004.

[4] Matthew A.J. Apps and Manos Tsakiris. The free-energy self: A predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, 41:85–97, 2014.

[5] M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, feb 2002.

[6] Frédéric Assal, Sophie Schwartz, and Patrik Vuilleumier. Moving with or without will: Functional neural correlates of alien hand syndrome. *Annals of Neurology*, 62(3):301–306, sep 2007.

[7] Per Bak and Maya Paczuski. Complexity, contingency, and criticality. *Proceedings of the National Academy of Sciences of the United States of America*, 92(15):6689–6696, jul 1995.

[8] H. B. Barlow. Possible Principles Underlying the Transformations of Sensory Messages. In *Sensory Communication*, pages 216–234. 1961.

[9] Danielle S. Bassett, Edward Bullmore, Beth A. Verchinski, Venkata S. Mattay, Daniel R. Weinberger, and Andreas Meyer-Lindenberg. Hierarchical organization of human cortical networks in health and schizophrenia. *Journal of Neuroscience*, 28(37):9239–9248, sep 2008.

[10] Stephen Batchelor. Greek Buddha: Pyrrho's encounter with early Buddhism in central Asia, 2016.

[11] Tim Bayne. *The Unity of Consciousness*. 2010.

[12] Ernest Becker. *The Denial of Death*. 1973.

[13] John M. Beggs. The criticality hypothesis: How local cortical networks might optimize information processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1864):329–343, feb 2008.

[14] Marlene Behrmann and Ruth Kimchi. What Does Visual Agnosia Tell Us About Perceptual Organization and Its Relationship to Object Perception? *Journal of Experimental Psychology: Human Perception and Performance*, 29(1):19–42, 2003.

[15] Marlene Behrmann and Mayu Nishimura. Agnosias. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(2):203–213, mar 2010.

[16] Jean François Billeter. *Études sur Tchouang-tseu*. Allia edition, 2004.

[17] Iftah Biran, Tania Giovannetti, Laurel Buxbaum, and Anjan Chatterjee. The alien hand syndrome: What makes the alien hand alien? *Cognitive Neuropsychology*, 23(4):563–582, jun 2006.

[18] Christopher Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1st ed. 20 edition, 2007.

[19] John Mark Bishop. A cognitive computation fallacy? Cognition, computations and panpsychism. *Cognitive Computation*, 1(3):221–233, may 2009.

[20] Scott R. Bishop, Mark Lau, Shauna Shapiro, Linda Carlson, Nicole D. Anderson, James Carmody, Zindel V. Segal, Susan Abbey, Michael Speca, Drew Velting, and Gerald Devins. Mindfulness: A proposed operational definition. *Clinical Psychology: Science and Practice*, 11(3):230–241, 2004.

[21] S J Blakemore, D M Wolpert, and C D Frith. Central cancellation of self-produced tickle sensation. *Nature neuroscience*, 1(7):635–640, 1998.

[22] Sarah J. Blakemore, Chris D. Frith, and Daniel M. Wolpert. Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience*, 11(5):551–559, 1999.

[23] Sarah Jayne Blakemore, Daniel M. Wolpert, and Christopher D. Frith. Abnormalities in the awareness of action. *Trends in Cognitive Sciences*, 6(6):237–242, 2002.

[24] Paul Blanck, Sarah Perleth, Thomas Heidenreich, Paula Kröger, Beate Ditzen, Hinrich Bents, and Johannes Mander. Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: Systematic review and meta-analysis. *Behaviour Research and Therapy*, 102:25–35, mar 2018.

[25] James C. Bliss, Hewitt D. Crane, Phyllis K. Mansfield, and James T. Townsend. Information available in brief tactile presentations. NASA CR-623. *NASA contractor report. NASA CR. United States. National Aeronautics and Space Administration*, 1(4):57–86, jul 1967.

[26] Ned Block. Two neural correlates of consciousness. *Trends in Cognitive Sciences*, 9(2):46–52, feb 2005.

[27] Bhikkhu Bodhi. *A Comprehensive Manual of Abhidhamma*. BPS Pariyatti Publishing, 2007.

[28] Istvan Bodnar. Aristotle's Natural Philosophy, may 2018.

[29] Juan P. Borda and Louis A. Sass. Phenomenology and neurobiology of self disorder in schizophrenia: Primary factors. *Schizophrenia Research*, 169(1-3):464–473, dec 2015.

[30] Peter Bossaerts and Carsten Murawski. From behavioural economics to neuroeconomics to decision neuroscience: The ascent of biology in research on human decision making. *Current Opinion in Behavioral Sciences*, 5:37–42, oct 2015.

[31] M M Botvinick, T S Braver, C S Carter, D M Barch, and J D Cohen. Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652, 2001.

[32] Matthew Botvinick and Marc Toussaint. Planning as inference. *Trends in Cognitive Sciences*, 16(10):485–488, oct 2012.

[33] Matthew M. Botvinick, Yael Niv, and Andrew C. Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, dec 2009.

[34] Randy L. Buckner and Daniel C. Carroll. Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2):49–57, feb 2007.

[35] Ed Bullmore and Olaf Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349, may 2012.

[36] Alex Byrne. Inverted Qualia. *Stanford Encyclopedia of Philosophy*, (143010005), 2008.

[37] Heather A Cameron and Lucas R Glover. Adult neurogenesis: beyond learning and memory. *Annual review of psychology*, 66:53–81, jan 2015.

[38] Peter Carruthers. How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32(02):121–138, 2009.

[39] B Carter. The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 310(1512):347–363, dec 1983.

[40] Andrea E Cavanna and Michael R Trimble. The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3):564–583, 2006.

[41] David J. Chalmers. Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3):200–219, sep 1995.

[42] David J. Chalmers. Does a rock implement every finite-state automaton? *Synthese*, 108(3):309–333, 1996.

[43] David J Chalmers, Richard Brown, Catarina Dutilh Novaes, Keith Frankish, Nick Humphrey, François Kammerer, Jackson Kernion, Uriah Kriegel, Tom Mcclelland, Luke Mcgowan, Kelvin Mcqueen, Luke Muehlhauser, Josh Weisberg, and David Yates. The Meta-Problem of Consciousness. *Journal of Consciousness Studies*, 25(9):6–61, 2018.

[44] Lucie Charles, Filip Van Opstal, Sébastien Marti, and Stanislas Dehaene. Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, 73:80–94, 2013.

[45] Yuhan Chen, Shengjun Wang, Claus C. Hilgetag, and Changsong Zhou. Trade-off between Multiple Constraints Enables Simultaneous Formation of Modules and Hubs in Neural Systems. *PLoS Computational Biology*, 9(3):e1002937, 2013.

[46] Dante R. Chialvo. Psychophysics: Are our senses critical? *Nature Physics*, 2(5):301–302, may 2006.

[47] Kalina Christoff, Diego Cosmelli, Dorothée Legrand, and Evan Thompson. Specifying the self for cognitive neuroscience. *Trends in Cognitive Sciences*, 15(3):104–112, mar 2011.

[48] Axel Cleeremans, Dalila Achoui, Arnaud Beauny, Lars Keuninckx, Jean Remy Martin, Santiago Muñoz-Moldes, Laurène Vuillaume, and Adélaïde de Heering. Learning to Be Conscious, feb 2020.

[49] Michael A Cohen, Daniel C Dennett, and Nancy Kanwisher. What is the Bandwidth of Perceptual Experience?, may 2016.

[50] Roger C. Conant and W. Ross Ashby. Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97, 1970.

[51] Hans Crauel, Arnaud Debussche, and Franco Flandoli. Random attractors. *Journal of Dynamics and Differential Equations*, 9(2):307–341, 1997.

[52] Francis Crick and Graeme Mitchison. The function of dream sleep. *Nature*, 304(5922):111–114, 1983.

[53] Antonio R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Penguin (Non-Classics), sep 1994.

[54] Charles Darwin. *On the Origin of Species*. Routledge, jun 1859.

[55] Christopher J. Darwin, Michael T. Turvey, and Robert G. Crowder. An auditory analogue of the sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology*, 3(2):255–267, apr 1972.

[56] Subimal Datta, Donald F. Siwek, Elissa H. Patterson, and Patsy B. Cipolloni. Localization of pontine PGO wave generation sites and their anatomical projections in the rat. *Synapse*, 30(4):409–423, 1998.

[57] J. Daunizeau, K.J. Friston, and S.J. Kiebel. Variational Bayesian identification and prediction of stochastic nonlinear dynamic causal models. *Physica D: Nonlinear Phenomena*, 238(21):2089–2118, nov 2009.

[58] Anthony S. David, Nicholas Bedford, Ben Wiffen, and James Gilleen. Failures of metacognition and lack of insight in neuropsychiatric disorders. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1379–1390, 2012.

[59] Martin Davis. Is mathematical insight algorithmic? *Behavioral and Brain Sciences*, 13(4):659–660, 1990.

[60] Peter Dayan and L. F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 1st edition, dec 2001.

[61] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The Helmholtz Machine. *Neural Computation*, 7(5):889–904, 1995.

[62] G C DeAngelis, I Ohzawa, and R D Freeman. Receptive-field dynamics in the central visual pathways. *Trends in neurosciences*, 18(10):451–458, 1995.

[63] Gustavo Deco, Edmund T. Rolls, and Ranulfo Romo. Stochastic dynamics as a principle of brain function. *Progress in Neurobiology*, 88(1):1–16, 2009.

[64] Javier DeFelipe, Henry Markram, and Kathleen S. Rockland. The neocortical column, jun 2012.

[65] S Dehaene and L Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79(1-2):1–37, 2001.

[66] S Dehaene, L Naccache, L Cohen, DL Bihan, JF Mangin, JB Poline, and D Rivière. Cerebral mechanisms of word masking and unconscious repetition priming. *Nature neuroscience*, 4(7):752–758, jul 2001.

[67] Stanislas Dehaene and Jean-Pierre Changeux. Experimental and Theoretical Approaches to Conscious Processing. *Neuron*, 70(2):200–227, apr 2011.

[68] Stanislas Dehaene, Jean-Pierre P Changeux, Lionel Naccache, Jérôme Sackur, and Claire Sergent. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, 10(5):204–211, may 2006.

[69] Stanislas Dehaene, Hakwan Lau, and Sid Kouider. What is consciousness, and could machines have it? *Science*, 358(6362):486–492, 2017.

[70] Stanislas Dehaene, Manuela Piazza, Philippe Pinel, and Laurent Cohen. Three parietal circuits for number processing. *Cognitive neuropsychology*, 20(3):487–506, 2003.

[71] Stanislas Dehaene, Claire Sergent, and Jean-Pierre P Changeux. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8520–8525, 2003.

[72] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[73] Daniel C. Dennett. Quining qualia. Oxford University Press, 1988.

[74] Daniel C. Dennett. Welcome to strong illusionism. *Journal of Consciousness Studies*, 26(9-10):48–58, 2019.

[75] René Descartes. *Discourse on method and Meditations on first philosophy.* Hackett Pub, 1641.

[76] R. Desimone, T. D. Albright, C. G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, 4(8):2051–2062, 1984.

[77] Susanne Diekelmann and Jan Born. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, jan 2010.

[78] Zoltan Dienes and Anil K. Seth. Measuring any conscious content versus measuring the relevant conscious content: Comment on Sandberg et al. *Consciousness and Cognition*, 19(4):1079–1080, apr 2010.

[79] Hugh Everett, John Archibald Wheeler, Bryce S. DeWitt, L. N. Cooper, D. Van Vechten, and Neill Graham. The many-worlds interpretation of quantum mechanics. pages 1–264, 1973.

[80] I. Feinberg. Efference copy and corollary discharge: implications for thinking and its disorders. *Schizophrenia Bulletin*, 4(4):636–640, 1978.

[81] Harriet Feldman and Karl J. Friston. Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, 4, 2010.

[82] D J Felleman and D C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, 1(1):1–47, jan 1991.

[83] Fernanda Ferreira, Karl G.D. Bailey, and Vittoria Ferraro. Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1):11–15, feb 2002.

[84] Richard Feynman, Leighton, and Sands. *The Feynman Lectures on Physics: Mainly mechanics, radiation, and heat.* 1963.

[85] Stephen M. Fleming. Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, 2020(1), jan 2020.

[86] Stephen M Fleming and Raymond J Dolan. The neural basis of metacognitive ability. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1594):1338–1349, 2012.

[87] Stephen M Fleming, Raymond J Dolan, and Christopher D Frith. Metacognition: computation, biology and function. *Philos Trans R Soc Lond B Biol Sci*, 367:1280+, 2012.

[88] Stephen M. Fleming, Jihye Ryu, John G. Golfinos, and Karen E. Blackmon. Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, 137(10):2811–2822, oct 2014.

[89] Paul C. Fletcher and Chris D. Frith. Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews Neuroscience*, 10(1):48–58, jan 2009.

[90] Alex Fornito, Andrew Zalesky, Christos Pantelis, and Edward T. Bullmore. Schizophrenia, neuroimaging and connectomics. *NeuroImage*, 62(4):2296–2314, oct 2012.

[91] K. J. Friston. Variational filtering. *NeuroImage*, 41(3):747–766, jul 2008.

[92] K J Friston and C D Frith. Schizophrenia: a disconnection syndrome? *Clinical neuroscience (New York, N.Y.)*, 3(2):89–97, jan 1995.

[93] K. J. Friston, C. D. Frith, R. E. Passingham, R. J. Dolan, P. F. Liddle, and R. S.J. Frackowiak. Entropy and cortical activity: information theory and PET findings. *Cerebral Cortex*, 2(3):259–67, 1992.

[94] K. J. Friston, N. Trujillo-Barreto, and J. Daunizeau. DEM: A variational treatment of dynamic systems. *NeuroImage*, 41(3):849–885, jul 2008.

[95] Karl Friston. Learning and inference in the brain. *Neural Networks*, 16(9):1325–1352, 2003.

[96] Karl Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836, 2005.

[97] Karl Friston. A free energy principle for a particular physics. jun 2019.

[98] Karl Friston, Michael Breakspear, and Gustavo Deco. Perception and self-organized instability. *Frontiers in Computational Neuroscience*, 6(JUL), jul 2012.

[99] Karl Friston, Harriet R. Brown, Jakob Siemerkus, and Klaas E. Stephan. The dysconnection hypothesis (2016). *Schizophrenia Research*, 176(2-3):83–94, oct 2016.

[100] Karl Friston and Christopher Frith. A Duet for one. *Consciousness and Cognition*, 36:390–405, nov 2015.

[101] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. *Journal of Physiology Paris*, 100(1-3):70–87, 2006.

[102] Karl Friston, Rosalyn Moran, and Anil K. Seth. Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology*, 23(2):172–178, 2013.

[103] Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6(4):187–214, 2015.

[104] Karl J Friston. Am I Self-Conscious? (Or Does Self-Organization Entail Self-Consciousness?). *Frontiers in Psychology*, 9:579+, 2018.

[105] Karl J. Friston, Jean Daunizeau, James Kilner, and Stefan J. Kiebel. Action and behavior: A free-energy formulation. *Biological Cybernetics*, 102(3):227–260, mar 2010.

[106] Karl J Friston, Jérémie Mattout, Nelson Trujillo-Barreto, John Ashburner, and Will Penny. Variational free energy and the Laplace approximation. *NeuroImage*, 34(1):220–234, 2007.

[107] Karl J. Friston, G. Tononi, O. Sporns, and G. M. Edelman. Characterising the complexity of neuronal interactions. *Human Brain Mapping*, 3(4):302–314, jan 1995.

[108] Karl J. Friston, Wanja Wiese, and J. Allan Hobson. Sentience and the origins of consciousness: From cartesian duality to Markovian monism. *Entropy*, 22(5):516, apr 2020.

[109] C. D. Frith. The positive and negative symptoms of schizophrenia reflect impairments in the perception and initiation of action. *Psychological Medicine*, 17(3):631–648, aug 1987.

[110] Christopher D. Frith, Sarah Jayne Blakemore, and Daniel M. Wolpert. Abnormalities in the Awareness and Control of Action. *Philosophical Transactions: Biological Sciences*, 355(1404):1771–1788, 2000.

[111] Christopher Donald Frith. *The cognitive neuropsychology of schizophrenia*. Erlbaum, feb 1992.

[112] Shaun Gallagher. Philosophical conceptions of the self: Implications for cognitive science. *Trends in Cognitive Sciences*, 4(1):14–21, jan 2000.

[113] Michael S. Gazzaniga. The Split Brain in Man. *Scientific American*, 217(2):24–29, aug 1967.

[114] Wulfram Gerstner and Werner M. Kistler. *Spiking Neuron Models*. Cambridge University Press, aug 2002.

[115] Seth J. Gillihan and Martha J. Farah. Is self special? A critical review of evidence from experimental psychology and cognitive neuroscience. *Psychological Bulletin*, 131(1):76–97, jan 2005.

[116] Philipp Goff, William Seager, and Sean Allen-Hermanson. Panpsychism, 2022.

[117] Joshua I. Gold and Michael N. Shadlen. The Neural Basis of Decision Making. *Annual Review of Neuroscience*, 30(1):535–574, jul 2007.

[118] Nigel Goldenfeld. *Lectures on Phase Transitions and the Renormalization Group*. CRC Press, mar 2018.

[119] Michael S. A. Graziano and Taylor W. Webb. The attention schema theory: a mechanistic account of subjective awareness. *Frontiers in Psychology*, 06(APR):23, apr 2015.

[120] David M Green and John A Swets. *Signal detection theory and psychophysics*. Wiley, New York, 1966.

[121] R. L. Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 290:181–197, 1980.

[122] Britta Hahn, Frank A. Wolkenberg, Thomas J. Ross, Carol S. Myers, Stephen J. Heishman, Dan J. Stein, Pradeep K. Kurup, and Elliot A. Stein. Divided versus selective attention: Evidence for common processing mechanisms. *Brain Research*, 1215:137–146, jun 2008.

[123] Hermann Haken. *Synergetics*, volume 1 of *Springer Series in Synergetics*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[124] Stuart Hameroff and Roger Penrose. Consciousness in the universe: A review of the 'Orch OR' theory, mar 2014.

[125] Deborah E Hannula, Daniel J Simons, and Neal J Cohen. Imaging implicit perception: promise and pitfalls. *Nature reviews. Neuroscience*, 6(3):247–255, mar 2005.

[126] Andrea Hasenstaub, Stephani Otte, Edward Callaway, and Terrence J. Sejnowski. Metabolic cost as a unifying principle governing neuronal biophysics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(27):12329–12334, jul 2010.

[127] Andrew M Haun, Giulio Tononi, Christof Koch, and Naotsugu Tsuchiya. Are we underestimating the richness of visual experience? *Neuroscience of Consciousness*, 2017(1), jan 2017.

[128] Jeff Hawkins and Sandra Blakeslee. *On Intelligence*. McMillan, 2004.

[129] H Von Helmholtz. *Handbuch der Physiologischen Optik*. 1867.

[130] M. G. Henriksen and J. Parnas. Self-disorders and Schizophrenia: A Phenomenological Reappraisal of Poor Insight and Noncompliance. *Schizophrenia Bulletin*, 40(3):542–547, may 2014.

[131] Hal E. Hershfield. The self over time. *Current Opinion in Psychology*, 26:72–75, apr 2019.

[132] Michael H Herzog, Thomas Kammer, and Frank Scharnowski. Time Slices: What Is the Duration of a Percept? *PLoS Biol*, 14(4):e1002433+, 2016.

[133] Jorge Hidalgo, Jacopo Grilli, Samir Suweis, Miguel A. Muñoz, Jayanth R. Banavar, and Amos Maritan. Information-based fitness and the emergence of criticality in living systems. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28):10095–10100, jul 2014.

[134] Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Radford M. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, may 1995.

[135] Geoffrey E. Hinton and Richard S Zemel. Autoencoders, Minimum Description Length and Helmholtz free Energy. *Advances in neural information processing systems*, 6:3–10, 1994.

[136] J. A. Hobson and K. J. Friston. Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, 98(1):82–98, jul 2012.

[137] J. A. Hobson and R. W. McCarley. The brain as a dream state generator: an activation-synthesis hypothesis of the dream process. *American Journal of Psychiatry*, 134(12):1335–1348, apr 1977.

[138] J. A. Hobson, E. F. Pace-Schott, and R. Stickgold. Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behavioral and Brain Sciences*, 23(6):793–842, 2000.

[139] J. Allan Hobson. REM sleep and dreaming: towards a theory of protoconsciousness. *Nature Reviews Neuroscience*, 10(11):803–813, nov 2009.

[140] Hohwy J and Michael J. Why Should Any Body Have a Self? In *The Subject's Matter*. The MIT Press, 2017.

[141] Peter C. Holland. Relations Between Pavlovian-Instrumental Transfer and Reinforcer Devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, 30(2):104–117, 2004.

[142] Charles Chong Hwa Hong, James C. Harris, Godfrey D. Pearlson, Jin Suh Kim, Vince D. Calhoun, James H. Fallon, Xavier Golay, Joseph S. Gillen, Daniel J. Simmonds, Peter C.M. Van Zijl, David S. Zee, and James J. Pekar. FMRI evidence for multisensory recruitment associated with rapid eye movements during sleep. *Human Brain Mapping*, 30(5):1705–1722, may 2009.

[143] Clark L. Hull. *A behavior system: an introduction to behavior theory concerning the individual organism*. Yale Press, 1953.

[144] David Hume. *A Treatise of Human Nature*. Clarendon Press, 1739.

[145] Barbara Jachs, Manuel J. Blanco, Sarah Grantham-Hill, and David Soto. On the independence of visual awareness and metacognition: A signal detection theoretic analysis. *Journal of Experimental Psychology: Human Perception and Performance*, 41(2):269–276, 2015.

[146] Frank Jackson. Epiphenomenal Qualia. *The Philosophical Quarterly*, 32(127):127, apr 1982.

[147] Philip L. Jackson and Jean Decety. Motor cognition: A new paradigm to study self-other interactions. *Current Opinion in Neurobiology*, 14(2):259–263, apr 2004.

[148] Gerald H. Jacobs. Evolution of colour vision in mammals, oct 2009.

[149] William James. *The Principles Of Psychology*. Number 1890. 1890.

[150] Gabriele Jordan, Samir S. Deeb, Jenny M. Bosten, and J. D. Mollon. The dimensionality of color vision in carriers of anomalous trichromacy. *Journal of Vision*, 10(8):12–12, jul 2010.

[151] Jon Kabat-Zinn. *Wherever you go, there you are*. Hyperion, New York, 1994.

[152] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, Transactions of the ASME*, 82(1):35–45, mar 1960.

[153] Ryota Kanai, Yutaka Komura, Stewart Shipp, and Karl Friston. Cerebral hierarchies: Predictive processing, precision and the pulvinar. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 2015.

[154] Immanuel Kant. *Critique of Pure Reason*. Palgrave Macmillan UK, London, 1787.

[155] Nandini Karunamuni and Rasanjala Weerasekera. Theoretical Foundations to Guide Mindfulness Meditation: A Path to Wisdom. *Current Psychology*, 38(3):627–646, jun 2019.

[156] Robert E. Kass and Duane Steffey. Approximate bayesian inference in conditionally independent hierarchical models (Parametric empirical bayes models). *Journal of the American Statistical Association*, 84(407):717–726, 1989.

[157] John Anderson Kay. *The truth about markets — Why Some Nations Are Rich but Most Remain Poor*. Allen Lane, 2003.

[158] Mehdi Keramati, Peter Smittenaar, Raymond J. Dolan, and Peter Dayan. Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. *Proceedings of the National Academy of Sciences of the United States of America*, 113(45):12868–12873, nov 2016.

[159] Stefan J. Kiebel, Jean Daunizeau, and Karl J. Friston. A Hierarchy of Time-Scales and the Brain. *PLoS Computational Biology*, 4(11):e1000209+, nov 2008.

[160] Robert Kirk. Sentience and behaviour. *Mind*, 83(329):43–60, jan 1974.

[161] Robert Kirk. Zombies, 2009.

[162] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, may 1983.

[163] D Knill and A Pouget. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends in neurosciences*, 27(12):712–719, 2004.

[164] Christof Koch and Naotsugu Tsuchiya. Attention and consciousness: two distinct brain processes. *Trends in Cognitive Sciences*, 11(1):16–22, jan 2007.

[165] Christopher F Kolb and Jochen Braun. Blindsight in normal observers. *Nature*, 377(6547):336–338, 1995.

[166] Sid Kouider, Vincent de Gardelle, Jérôme Sackur, and Emmanuel Dupoux. How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences*, 14(7):301–307, jun 2010.

[167] Peter D. Kvam, Timothy J. Pleskac, Shuli Yu, and Jerome R. Busemeyer. Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences of the United States of America*, 112(34):10645–10650, aug 2015.

[168] B. Laeng, S. M. Kosslyn, V. S. Caviness, and J. Bates. Can deficits in spatial indexing contribute to simultanagnosia? *Cognitive Neuropsychology*, 16(2):81–114, mar 1999.

[169] Diogenes Laertius. *Lives of Eminent Philosophers*. Cambridge University Press, may 200.

[170] V A Lamme and P R Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, 2000.

[171] Victor A Lamme. Why visual attention and awareness are different. *Trends in Cognitive Sciences*, 7(1):12–18, jan 2003.

[172] Victor A Lamme. Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, 10(11):494–501, nov 2006.

[173] Laozi. *Lao zi (Dao De Jing)*. Ediciones Alfaguara.

[174] Hakwan Lau. Consciousness, Metacognition, and Perceptual Reality Monitoring. *PsyArXiv*, pages 1–17, 2019.

[175] Hakwan Lau and David Rosenthal. Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8):365–373, aug 2011.

[176] Hakwan C. Lau and Richard E. Passingham. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proceedings of the National Academy of Sciences*, 103(49):18763–18768, dec 2006.

[177] Steven Laureys, Adrian M. Owen, and Nicholas D. Schiff. Brain function in coma, vegetative state, and related disorders. *Lancet Neurology*, 3(9):537–546, sep 2004.

[178] Dorothée Legrand and Perrine Ruby. What Is Self-Specific? Theoretical Investigation and Critical Review of Neuroimaging Results. *Psychological Review*, 116(1):252–282, 2009.

[179] John C Lilly. *The Scientist: A Metaphysical Autobiography*. 1996.

[180] R Linsker. Perceptual Neural Organization: Some Approaches Based on Network Models and Information Theory. *Annual Review of Neuroscience*, 13(1):257–281, mar 1990.

[181] John Locke. *An essay concerning human understanding*. J.M. Dent ;;Dutton, London ;New York, 1690.

[182] George Loewenstein. Preferences, behavior, and welfare: Emotions in economic theory and economic behavior. *American Economic Review*, 90(2):426–432, may 2000.

[183] Hans C. Lou, Bruce Luber, Michael Crupain, Julian P. Keenan, Markus Nowak, Troels W. Kjaer, Harold A. Sackeim, and Sarah H. Lisanby. Parietal cortex and representation of the mental Self. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6827–6832, apr 2004.

[184] J R Lucas. Minds, Machines and Gödel. *Philosophy*, XXXVI:112–127, 1961.

[185] Long Luu and Alan A Stocker. Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife*, 7:e33334, 2018.

[186] Steven A. Marchette, Lindsay K. Vass, Jack Ryan, and Russell A. Epstein. Anchoring the neural compass: Coding of local spatial reference frames in human medial parietal lobe. *Nature Neuroscience*, 17(11):1598–1606, oct 2014.

[187] Ivana S. Marková and German E. Berrios. The construction of anosognosia: History and implications. *Cortex*, 61:9–17, dec 2014.

[188] D. C. Marr and T. Poggio. From understanding computation to understanding neural circuitry. *Neurosciences Research Program Bulletin*, 15(3):470–488, may 1977.

[189] Brice Martin, Marc Wittmann, Nicolas Franck, Michel Cermolacce, Fabrice Berna, and Anne Giersch. Temporal structure of consciousness and minimal self in schizophrenia. *Frontiers in Psychology*, 5:1175, oct 2014.

[190] Mario Martinez-Saito. A Skeptical View on the Physics-Consciousness Explanatory Gap. *Axiomathes*, 32(6):1081–1110, dec 2022.

[191] Mario Martinez-Saito. Discrete scaling and criticality in a chain of adaptive excitable integrators. *Chaos, Solitons and Fractals*, 163:112574, oct 2022.

[192] Mario Martinez-Saito. Probing doors to visual awareness: Choice set, visibility, and confidence. *Visual Cognition*, 30(6):393–424, jul 2022.

[193] George A. Mashour, Pieter Roelfsema, Jean Pierre Changeux, and Stanislas Dehaene. Conscious Processing and the Global Neuronal Workspace Hypothesis, mar 2020.

[194] Iacopo Mastromatteo and Matteo Marsili. On the criticality of inferred models. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(10):P10012, oct 2011.

[195] Tim Maudlin. Computation and Consciousness. *The Journal of Philosophy*, 86(8):407, aug 1989.

[196] M.-Marsel Mesulam. From sensation to cognition. *Brain*, 121:1013–1052, 1998.

[197] Thomas Metzinger. *Being no one: The self-model theory of subjectivity.* MIT Press, 2003.

[198] Thomas Metzinger. The myth of cognitive agency: Subpersonal thinking as a cyclically recurring loss of mental autonomy. *Frontiers in Psychology*, 4(DEC):931, 2013.

[199] Florent Meyniel, Daniel Schlunegger, and Stanislas Dehaene. The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLoS Computational Biology*, 11(6), jun 2015.

[200] Vernon B. Mountcastle. Modality and topographic properties of single neurons of cat's somatic sensory cortex. *Journal of neurophysiology*, 20(4):408–434, jul 1957.

[201] D Mumford. On the computational architecture of the neocortex. *Biological Cybernetics*, 65(2):135–145, 1991.

[202] D. Mumford. On the computational architecture of the neocortex - II The role of cortico-cortical loops. *Biological Cybernetics*, 66(3):241–251, jan 1992.

[203] Thomas Nagel. What Is It Like to Be a Bat? *The Philosophical Review*, 83(4):435–450, oct 1974.

[204] Antonio Narzisi and Rosy Muccio. A Neuro-Phenomenological Perspective on the Autism Phenotype. *Brain Sciences*, 11(7):914, jul 2021.

[205] Ulric Neisser. Five kinds of self-knowledge. *Philosophical Psychology*, 1(1):35–59, jan 1988.

[206] Friedrich Nietzsche. *The Joyful Wisdom*. Project Gutenberg, 1882.

[207] Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, mar 1977.

[208] Tor Nørretranders. *The user illusion: cutting consciousness down to size*. New York: Viking, 1991.

[209] Brian Odegaard, Min Yu Chang, Hakwan Lau, and Sing Hang Cheung. Inflation versus filling-in: Why we feel we see more than we actually do in peripheral vision. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1755), sep 2018.

[210] Brian Odegaard, Piercesare Grimaldi, Seong H Cho, Megan A K Peters, Hakwan C Lau, and Michele A Basso. Superior colliculus neuronal ensemble activity signals optimal rather than subjective confidence. *Proceedings of the National Academy of Sciences of the United States of America*, 115(7):E1588—-E1597, 2018.

[211] Zeev Olami, Hans Jacob S. Feder, and Kim Christensen. Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Physical Review Letters*, 68(8):1244–1247, feb 1992.

[212] Morten Overgaard, Julian Rote, Kim Mouridsen, and Thomas Zoëga Ramsøy. Is conscious perception gradual or dichotomous? A comparison of report methodologies during a visual task. *Consciousness and Cognition*, 15(4):700–708, dec 2006.

[213] Morten Overgaard and Kristian Sandberg. Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1287–1296, 2012.

[214] Josef Parnas, Paul Møller, Tilo Kircher, Jørgen Thalbitzer, Lennart Jansson, Peter Handest, and Dan Zahavi. EASE: Examination of Anomalous Self-Experience. *Psychopathology*, 38(5):236–58, sep 2005.

[215] Megan A.K. Peters and Hakwan Lau. Human observers have optimal introspective access to perceptual processes even for visually masked stimuli. *eLife*, 4(OCTOBER2015):e09651, oct 2015.

[216] Plato. *Diálogos I: Apología de Sócrates, Critón, Eutifrón, Hipias Menor, Hipias Mayor, Ion, Lisis, Cármides, Laques y Protágoras*. Gredos.

[217] N. Platt, E. A. Spiegel, and C. Tresser. On-off intermittency: A mechanism for bursting. *Physical Review Letters*, 70(3):279–282, jan 1993.

[218] L. Postmes, H. N. Sno, S. Goedhart, J. van der Stel, H. D. Heering, and L. de Haan. Schizophrenia as a self-disorder due to perceptual incoherence, jan 2014.

[219] Hilary Putnam. Mind, Language and Reality: Philosophical Papers, volume 2, 1975.

[220] Sepehrdad Rahimian. The myth of when and where: How false assumptions still haunt theories of consciousness. *Consciousness and Cognition*, 97:103246, jan 2022.

[221] V. S. Ramachandran and William Hirstein. The perception of phantom limbs. The D. O. Hebb lecture. *Brain*, 121(9):1603–1630, 1998.

[222] Maxwell James Désormeau Ramstead, Paul Benjamin Badcock, and Karl John Friston. Answering Schrödinger's question: A free-energy formulation. *Physics of Life Reviews*, 24:1–16, mar 2018.

[223] A. Revonsuo. The reinterpretation of dreams: An evolutionary hypothesis of the function of dreaming. *Behavioral and Brain Sciences*, 23(6):877–901, 2000.

[224] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, nov 1999.

[225] Marion Rouault, Maël Lebreton, and Mathias Pessiglione. A shared brain system forming confidence judgment across cognitive domains. *Cerebral Cortex*, 33(4):1426–1439, feb 2023.

[226] Perrine Ruby and Jean Decety. Effect of subjective perspective taking during simulation of action: A PET investigation of agency. *Nature Neuroscience*, 4(5):546–550, 2001.

[227] Gilbert Ryle. *The concept of mind*. University of Chicago Press, 1949.

[228] Oliver Sacks. *The Man Who Mistook His Wife For A Hat and other clinical tales*. Harpercollins, 1985.

[229] Bastian Sajonz, Thorsten Kahnt, Daniel S. Margulies, Soyoung Q. Park, André Wittmann, Meline Stoy, Andreas Ströhle, Andreas Heinz, Georg Northoff, and Felix Bermpohl. Delineating self-referential processing from episodic memory retrieval: Common and dissociable networks. *NeuroImage*, 50(4):1606–1617, may 2010.

[230] Elyn R. Saks. *The center cannot hold: my journey through madness*. Hachette Books, 2008.

[231] M. D. Sanders, Elizabeth K. Warrington, John Marshall, and L. Wieskrantz. "Blindsight": vision a field defect. *The Lancet*, 303(7860):707–708, apr 1974.

[232] L. A. Sass and J. Parnas. Schizophrenia, Consciousness, and the Self. *Schizophrenia Bulletin*, 29(3):427–444, jan 2003.

[233] Louis A. Sass. Self-disturbance and schizophrenia: Structure, specificity, pathogenesis (Current issues, New directions). *Schizophrenia Research*, 152(1):5–11, jan 2014.

[234] Louis A. Sass and Juan P. Borda. Phenomenology and neurobiology of self disorder in schizophrenia: Secondary factors. *Schizophrenia Research*, 169(1-3):474–482, dec 2015.

[235] Daniel L. Schacter. *The seven sins of memory: how the mind forgets and remembers*. Houghton Mifflin, 2001.

[236] Daniel L. Schacter, Donna Rose Addis, and Randy L. Buckner. Remembering the past to imagine the future: The prospective brain. *Nature Reviews Neuroscience*, 8(9):657–661, sep 2007.

[237] Frank Scharnowski, Johannes Rüter, Jacob Jolij, Frouke Hermens, Thomas Kammer, and Michael H. Herzog. Long-lasting modulation of feature integration by transcranial magnetic stimulation. *Journal of Vision*, 9(6):1–1, jun 2009.

[238] Taylor W. Schmitz and Sterling C. Johnson. Relevance to self: A brief review and framework of neural systems underlying appraisal. *Neuroscience and Biobehavioral Reviews*, 31(4):585–596, 2007.

[239] Jonathan W Schooler. Re-representing consciousness: dissociations between experience and meta-consciousness. *Trends in Cognitive Sciences*, 6(8):339–344, 2002.

[240] Arthur Schopenhauer. *On the Freedom of the Will*. 1839.

[241] Arthur Schopenhauer. *The World as Will and Representation, Vol. 1*. Dover Publications Inc., 1844.

[242] Arthur Schopenhauer. *The World as Will and Representation, Vol. 2*. Dover Publications Inc., 1844.

[243] Frauke Schultze-Lutter. Subjective symptoms of schizophrenia in research and the clinic: The basic symptom concept, jan 2009.

[244] Gregory Scott, Erik D. Fagerholm, Hiroki Mutoh, Robert Leech, David J. Sharp, Woodrow L. Shew, and Thomas Knöpfel. Voltage imaging of waking mouse cortex reveals emergence of critical neuronal dynamics. *Journal of Neuroscience*, 34(50):16611–16620, dec 2014.

[245] Claire Sergent, Sylvain Baillet, and Stanislas Dehaene. Timing of the brain events underlying access to consciousness during the attentional blink. *Nature Neuroscience*, 8(10):1391–1400, sep 2005.

[246] Claire Sergent and Stanislas Dehaene. Is consciousness a gradual phenomenon? Evidence for an all-or-none bifurcation during the attentional blink. *Psychological science : a journal of the American Psychological Society / APS*, 15(11):720–728, nov 2004.

[247] Claire Sergent, Valentin Wyart, Mariana Babo-Rebelo, Laurent Cohen, Lionel Naccache, and Catherine Tallon-Baudry. Cueing attention after the stimulus is gone can retrospectively trigger conscious perception. *Current Biology*, 23(2):150–155, 2013.

[248] Anil K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, nov 2013.

[249] Anil K. Seth. A predictive processing theory of sensorimotor contingencies: Explaining the puzzle of perceptual presence and its absence in synesthesia. *Cognitive Neuroscience*, 5(2):97–118, 2014.

[250] Anil K. Seth and Tim Bayne. Theories of consciousness. *Nature Reviews Neuroscience*, 23(7):439–452, jul 2022.

[251] Anil K Seth, Zoltán Dienes, Axel Cleeremans, Morten Overgaard, and Luiz Pessoa. Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in cognitive sciences*, 12(8):314–321, aug 2008.

[252] Anil K. Seth and Karl J. Friston. Active interoceptive inference and the emotional brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160007, nov 2016.

[253] Anil K. Seth, Keisuke Suzuki, and Hugo D. Critchley. An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, 3(JAN):395, 2012.

[254] Anil K Seth, Commentator Wanja Wiese, Johannes Gutenberg, Thomas Metzinger, and Jennifer M Windt. Inference to the Best Prediction. *Open MIND*, 35:1–8, 2015.

[255] Gordon M.G. Shepherd and Naoki Yamawaki. Untangling the cortico-thalamo-cortical loop: cellular pieces of a knotty circuit puzzle. *Nature Reviews Neuroscience 2021 22:7*, 22(7):389–406, may 2021.

[256] S. Murray Sherman. The thalamus is more than just a relay. *Current Opinion in Neurobiology*, 17(4):417–422, aug 2007.

[257] Woodrow L Shew and Dietmar Plenz. The functional benefits of criticality in the cortex. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 19(1):88–100, 2013.

[258] Sydney Shoemaker. *Identity, Cause, and Mind: Philiosophical Essays*. 2003.

[259] Camilo Miguel Signorelli, Quanlong Wang, and Ilyas Khan. A compositional model of consciousness based on consciousness-only. *Entropy*, 23(3):1–18, mar 2021.

[260] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, mar 1956.

[261] Jonathan Smallwood. Mind-wandering While Reading: Attentional Decoupling, Mindless Reading and the Cascade Model of Inattention. *Linguistics and Language Compass*, 5(2):63–77, feb 2011.

[262] Jonathan Smallwood and Jonathan W. Schooler. The science of mind wandering: Empirically navigating the stream of consciousness. *Annual Review of Psychology*, 66:487–518, jan 2015.

[263] M Snodgrass. Disambiguating conscious and unconscious influences: do exclusion paradigms demonstrate unconscious perception? *Am J Psychol*, 115(4):545–579, 2002.

[264] Michael Snodgrass, Edward Bernat, and Howard Shevrin. Unconscious perception: a model-based approach to method and evidence. *Perception and Psychophysics*, 66(5):846–67, jul 2004.

[265] M. Solms. Dreaming and REM sleep are controlled by different brain mechanisms. *Behavioral and Brain Sciences*, 23(6):843–850, 2000.

[266] Alec Solway and Matthew M. Botvinick. Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences of the United States of America*, 112(37):11708–11713, sep 2015.

[267] Roberto C. Sotero, Nelson J. Trujillo-Barreto, Yasser Iturria-Medina, Felix Carbonell, and Juan C. Jimenez. Realistically coupled neural mass models can generate EEG rhythms. *Neural Computation*, 19(2):478–512, 2007.

[268] Sean A. Spence. Free Will in the Light of Neuropsychiatry. *Philosophy, Psychiatry, and Psychology*, 3(2):75–90, 1996.

[269] G Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74:1–29, 1960.

[270] G. Sperling. Successive approximations to a model for short term memory. *Acta Psychologica*, 27(C):285–292, jan 1967.

[271] Alan A. Stocker and Eero P. Simoncelli. A Bayesian model of conditioned perception. *Advances in Neural Information Processing Systems 20*, 20:1409–1416, 2007.

[272] Galen Strawson. The self. *Journal of Consciousness Studies*, 4:405–428, 1997.

[273] Galen Strawson. Self, Body, and Experience. *Aristotelian Society Supplementary Volume*, 73(1):307–332, jul 1999.

[274] Galen Strawson. The self and the SESMET. *Journal of Consciousness Studies*, 6(4):99–135, apr 1999.

[275] Galen Strawson. Realistic Monism: Why Physicalism Entails Panpsychism. *Journal of Consciousness Studies*, 13(10-11):3–31, may 2006.

[276] Galen Strawson. A hundred years of consciousness: "A long training in absurdity". *Estudios de Filosofía*, (59):9–43, jan 2019.

[277] S M Stringer, T P Trappenberg, E T Rolls, and I E de Araujo. Self-organizing continuous attractor networks and path integration: one-dimensional models of head direction cells. *Network*, 13(2):217–242, 2002.

[278] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[279] Jun Tani. An interpretation of the "self" from the dynamical systems perspective: A constructivist approach. *Journal of Consciousness Studies*, 5:516–542, 1998.

[280] Max Tegmark. Why the brain is probably not a quantum computer. *Information sciences*, 128(3):155–179, oct 2000.

[281] Emanuel Todorov and Michael I. Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002.

[282] Giulio Tononi, Melanie Boly, Marcello Massimini, and Christof Koch. Integrated information theory: from consciousness to its physical substrate. *Nature reviews. Neuroscience*, 17(7):450–61, jul 2016.

[283] Giulio Tononi and Chiara Cirelli. Sleep function and synaptic homeostasis. *Sleep Medicine Reviews*, 10(1):49–62, feb 2006.

[284] Giulio Tononi, Olaf Sporns, and Gerald M. Edelman. A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, 91(11):5033–5037, may 1994.

[285] Thomas Trappenberg. *Fundamentals of Computational Neuroscience*. Oxford University Press, USA, 2002.

[286] Wolfgang Tschacher and Hermann Haken. Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas in Psychology*, 25(1):1–15, apr 2007.

[287] Naotsugu Tsuchiya, Melanie Wilke, Stefan Frässle, and Victor A.F. Lamme. No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences*, 19(12):757–770, 2015.

[288] Endel Tulving. Episodic memory: From mind to brain. *Annual Review of Psychology*, 53:1–25, 2002.

[289] Giuseppe Vallar and Roberta Ronchi. Anosognosia for motor and sensory deficits after unilateral brain damage: A review. *Restorative Neurology and Neuroscience*, 24(4-6):247–257, 2006.

[290] C van Vreeswijk and H Sompolinsky. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science*, 274(5293):1724–1726, dec 1996.

[291] F. G. Varela, H. R. Maturana, and R. Uribe. Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems*, 5(4):187–196, may 1974.

[292] Francisco Varela. The specious present: A neurophenomenology of time consciousness. In J. Petitot, F. J. Varela, and B. PachoudJ.-M. Roy, editors, *Naturalizing phenomenology: Issues in contemporary phenomenology and cognitive science*, pages 266–314. 1999.

[293] Francisco J Varela, Evan Thompson, and Eleanor Rosch. Enaction: Embodied. *The Embodied Mind: Cognitive Science and Human Experience*, 1991.

[294] Erich von Holst and Horst Mittelstaedt. Das Reafferenzprinzip - Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Die Naturwissenschaften*, 37(20):464–476, 1950.

[295] John von Neumann. *Mathematical Foundations of Quantum Mechanics (trans. from German by R. T. Beyer, 1955)*. Princeton University Press, 1955.

[296] Basil Wahn and Peter König. Is attentional resource allocation across sensory modalities task-dependent? *Advances in Cognitive Psychology*, 13(1):83–96, 2017.

[297] David J. Wales and Jonathan P.K. Doye. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *Journal of Physical Chemistry A*, 101(28):5111–5116, jul 1997.

[298] Garrett Wendel, Luis Martínez, and Martin Bojowald. Physical Implications of a Fundamental Period of Time. *Physical Review Letters*, 124(24):241301, jun 2020.

[299] Eugene Wigner and Henry Margenau. Symmetries and Reflections, Scientific Essays. *American Journal of Physics*, 35(12):1169–1170, jul 1967.

[300] Ludwig Wittgenstein. *The Blue and Brown Books*. Blackwell Oxford, 1958.

[301] D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, 11(7-8):1317–1329, 1998.

[302] Daniel M. Wolpert and Zoubin Ghahramani. Computational principles of movement neuroscience. *Nature Neuroscience*, 3(11s):1212–1217, 2000.

[303] Daniel M Wolpert, Zoubin Ghahramani, and Michael I Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, sep 1995.

[304] S Zeki and S Shipp. The functional logic of cortical connections. *Nature*, 335(6188):311–317, 1988.

[305] Semir Zeki. The disunity of consciousness, 2007.

[306] Zhuangzi. and Burton Watson. *Basic writings*. Columbia University Press, 1996.

[307] Ariel Zylberberg, Pablo Barttfeld, and Mariano Sigman. The construction of confidence in a perceptual decision. *Frontiers in Integrative Neuroscience*, 6, 2012.