Beyond Popular Science

David H. Silver

BEYOND POPULAR SCIENCE

David H. Silver

https://www.openbookpublishers.com

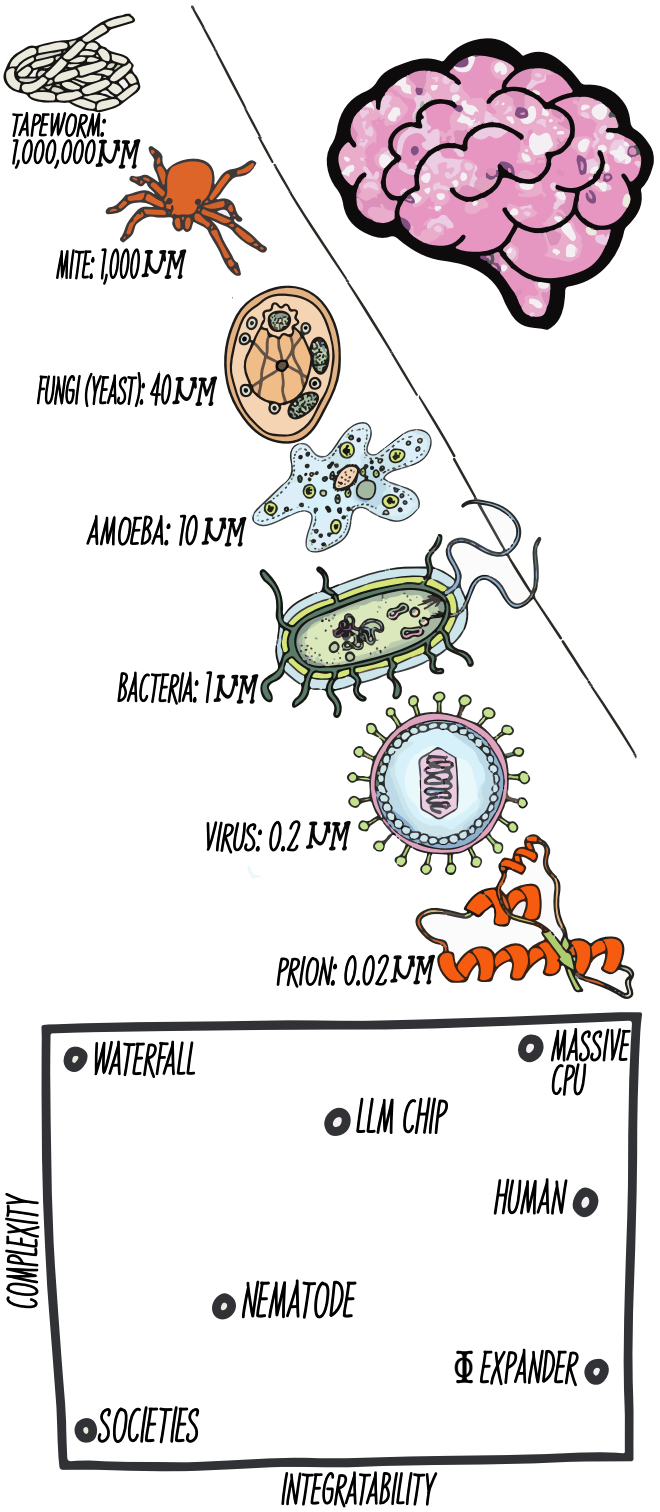
All external links were active at the time of publication unless otherwise stated and have been archived via the Internet Archive Wayback Machine at https://archive.org/web

Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0526#resources

ISBN Paperback:	978-1-80511-877-0
ISBN Hardback:	978-1-80511-878-7
ISBN Digital (PDF):	978-1-80511-879-4
ISBN HTML:	978-1-80511-881-7
ISBN Digital ebook (epub):	978-1-80511-880-0
DOI:	10.11647/OBP.0526

Cover image by Enny Silver and David H. Silver
Cover design by Jeevanjot Kaur Nagpal

A Freely Wilful Ignorance

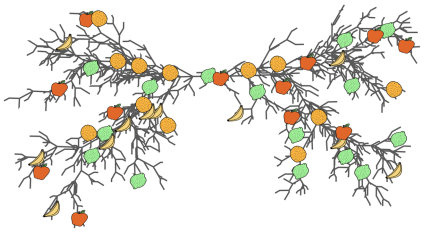


Top (Scales of Infection and Pathogenic Agents): Infectious agents span eight orders of magnitude—from macroscopic parasites such as tapeworms to sub-viral prions. Each has its own transmission pathway, lifecycle, and interaction mode with the host. Top-right: a brain, target of prion neurodegeneration.

Bottom (Post-Hoc Theories of Consciousness): Proposed mechanisms for consciousness map along a complexity–integrability plane: waterfall-like chaos, chip design, nematode circuits, human brains, and societal behaviour. But all are descriptive, not explanatory. No theory derives subjective experience from first principles—only correlates it post hoc to system properties, and for each example we can find synthetic constructs as a counter-example. For instance, if we make a chip composed of massively parallel XOR gates arranged in feedback loops to maximise cross-dependence, it can be tuned to produce an arbitrarily high φ value—far exceeding estimates for the human brain—yet the device is nothing more than a repetitive logical expander.

A Freely Wilful Ignorance

Milligrams of propofol erase consciousness in seconds. Fatal familial insomnia prevents its cessation for months until death. While we can reliably toggle awareness, no unified mechanism explains why subjectivity vanishes. Consciousness cannot be reduced to neural correlates or fit by classifiers. Any attempt to locate its origin in physical mechanisms presupposes the very phenomenon under study. Free will and physics appear incompatible, but the standoff is asymmetric: agency is the lived fact that makes physics construction possible. Consciousness occupies the apex of a revision hierarchy where, in any conflict with lower-level descriptions, the knower must prevail.



Consciousness & Anaesthesia ∘ General Anaesthetic Mystery ∘ Multiple Molecular Mechanisms ∘ Fatal Familial Insomnia ∘ Free Will vs Physics ∘ First-Person Experience ∘ Libet Readiness Potential ∘ Revision Cost Hierarchy ∘ Cogito Ergo Sum ∘ Direct Self-Knowledge ∘ Hard Problem

«الاعتقاد ليس هو المعنى المقصود، بل المعنى المتصور في النفس»

("Belief is not the utterance, but the conception in the soul.")
— Maimonides, circa 1191 CE

"Reason will prevail."
— The Gang, 2008

A Freely Wilful Ignorance

Ether-era anaesthesia began in 1846 with Morton's public demonstration in Boston; within months, ether and chloroform spread worldwide. By the early twentieth century, Meyer and Overton independently observed a correlation: anaesthetic potency scaled with lipid solubility across diverse compounds. This supported the idea that consciousness could be turned off by a nonspecific action on neuronal membranes. Yet the correlation cracked under scrutiny: highly lipophilic yet inert molecules failed to anaesthetise, while effective agents deviated from the predicted potency.

Mid-to-late twentieth century work shifted toward specific molecular targets: $GABA_A$ receptors, NMDA antagonism, two-pore K^+ channels, and hyperpolarisation-activated cyclic nucleotide–gated (HCN1) currents. Still, no single pathway unified the class.

In prion disease, a different historical thread exposed the opposite failure mode. In 1982, Prusiner proposed prions—a proteinaceous infectious particles—as agents of neurodegeneration. A rare PRNP mutation producing fatal familial insomnia (FFI) was later traced to selective thalamic degeneration, abolishing sleep despite otherwise preserved wakeful function. An Italian pedigree provided the defining clinical arc: onset with fragmented sleep, inexorable insomnia, autonomic failure, cognitive collapse, and death within months. Where anaesthesia induced obliviousness, FFI prevented it.

Reader beware: this chapter is not about a phenomenon at the heart of the scientific consensus; instead it is full of my philosophical musings.

General anaesthesia abolishes subjectivity itself. Other drugs alter perception, mood, or pain. Anaesthetics suspend the condition for all perception and mood. A standard intravenous dose of propofol—two milligrams per kilogram—eliminates awareness in less than a minute. The transition is sharp. One moment the subject tracks voices and surroundings; the next moment there is no report, no continuity of thought, and no subsequent memory. The effect is reliable, reversible, and indispensable to surgical practice. Yet it remains unexplained. That is assuming the loss is real, and not merely after the fact amnesia.

Different drugs converge on this endpoint through divergent and sometimes contradictory mechanisms. Propofol potentiates γ-aminobutyric acid type A ($GABA_A$) receptors, amplifying inhibitory currents and reducing excitability across the cortex. Isoflurane, sevoflurane, and other volatile anaesthetics bind to potassium and sodium channels, producing generalised dampening of neuronal firing. Nitrous oxide and xenon inhibit *N*-methyl-D-aspartate (NMDA) receptors, reducing excitatory drive. Ketamine blocks NMDA receptors yet increases cortical activity globally, producing electroencephalographic patterns closer to wakefulness than sleep while still abolishing awareness. Distinct molecular actions—some silencing neurons, some exciting them—dismantle consciousness with similar reliability.

The search for an underlying model for general anaesthesia once looked promising. At the turn of the twentieth century, Hans Meyer and Charles Ernest Overton noted a correlation: anaesthetic potency scales with lipid solubility. The Meyer–Overton rule suggested that

 https://doi.org/10.11647/OBP.0526.50

anaesthetics (Meyer, 1899; Overton, 1901) dissolved into neuronal membranes, altering their physical properties. For decades this correlation dominated, reinforced by its simplicity. Yet the correlation is not absolute. Non-immobilisers—molecules with high lipid solubility—fail to anaesthetize. Others deviate from predicted potency. The membrane theory could not account for exceptions.

The focus moved to receptors. Different anaesthetic classes bind to distinct proteins: $GABA_A$, NMDA, and two-pore domain potassium channels among prime candidates. Yet receptor theories also encounter anomalies. No single target is necessary. Mice engineered with $GABA_A$ subunits resistant to volatile anaesthetics still lose consciousness when exposed. No single target is sufficient: receptor agonists or antagonists with precise effects on candidate pathways often fail to produce general anaesthesia. What remains is a map of partial correlates, not a law specifying why awareness vanishes.

Network hypotheses move up a level. Thalamic "switch-off" models propose that sensory relay and intralaminar nuclei disengage cortical broadcasting. Alternatives hold that long-range cortico-cortical integration degrades: effective connectivity fragments, ignition-like reverberation collapses, and fronto-parietal synchrony decouples. Empirically, anaesthetic depth tracks changes in spectral power, complexity, and coherence. Though counterexamples persist. Ketamine increases cortical activity and high-frequency power yet abolishes consciousness. Dexmedetomidine reduces thalamic throughput yet permits vivid dreams.

The opposite extreme also exists. Infectious agents span orders of magnitude: from metre-long tapeworms to micrometre bacteria and nanometre viruses. Though some infections are not carried by biological agents, but by physical ones. A prion (proteinaceous infectious particle) is a protein (Prusiner, 1982) at nanometre scale) that was misfolded into an abnormal shape and can sometimes infect nearby proteins to do the same. It resists most disinfection protocols. And it can cause a consciousness disorder.

Fatal familial insomnia, a prion disease (Lugaresi et al., 1986), destroys neurons in the thalamus, especially in the anteroventral and mediodorsal nuclei. These nuclei regulate sleep architecture. As they degenerate, the subject loses the ability to enter non-rapid eye movement sleep. Ordinary fatigue accumulates, but sleep never arrives. Patients remain in escalating wakefulness until death, typically within one to two years of symptom onset. Consciousness persists compulsively until the body collapses under uninterrupted wakefulness.

Anaesthesia and prion disease bracket the same mystery. Milligrams of a synthetic molecule suspend awareness entirely. Widespread neuronal loss fails to interrupt it. Consciousness is too easy to subtract and, simultaneously, impossible to eliminate. This indicates that manipulations reach only the conditions under which consciousness manifests. They do not specify what consciousness is. Practitioners can toggle the switch without knowing what is being switched.

Measuring consciousness remains harder than turning it off. Clinical scales rely on responsiveness; neurophysiology adds proxies: cross-regional EEG (Electroencephalography) coherence, perturbational complexity from TMS-evoked (transcranial magnetic stimulation)

responses, and theoretical constructs such as Integrated Information Theory's Φ (Tononi, 2004). Each stumbles. Some unresponsive patients process speech. High Φ can be assigned to systems with no plausible subjectivity. EEG signatures of wakefulness can appear under amnestic sedation. Competing theories—Global Workspace, Integrated Information, Recurrent Processing—disagree on what makes a state conscious, and experiments often adjudicate proxies rather than experience itself.

The working picture is that multiple molecular routes converge on a few network-level motifs—reduced ignition, impaired integration, altered thalamocortical gating—sufficient to block access to a reportable workspace. That picture explains much of practice and little of essence.

The gap between control and understanding demands a different frame. Consciousness is singular. Treating it as a parameter vector to be fit by a classifier condescends to the phenomenon. A classifier extracts invariants and separates classes. Consciousness is first-personal presence and deliberative control. No change of basis, no loss function turns one into the other. The distinction is categorical.

Any research programme that seeks to locate the origin of consciousness in physical mechanisms presupposes the very phenomenon it attempts to explain. You deploy attention, select among hypotheses, compare results, and conclude. Each act exercises the thing under study. This reflexivity marks a boundary of intelligibility: the point where explanation reaches its natural terminus because the explanans and the explanandum coincide. Thomas Reid identified reflexive self-awareness as a first principle of common sense—an immediate, non-derivable truth that grounds all inquiry. Consciousness, when reflecting on itself, encounters not an epistemic obstacle but the foundational condition for explanation itself.

Free will and physics appear incompatible. If physics is a complete description—deterministic or stochastic, local or quantum, simulated or fundamental—then every decision reduces to a trajectory in state space. Free will becomes an illusion, a narrative that complex systems tell themselves about their own deterministic unfolding. But if free will exists, then physics is inconsistent. The standoff seems symmetric: pick your side.

The symmetry is false. Free will is the lived fact. Physics is the constructed model. If physics denies free will, physics has misclassified its own status. Constructing, testing, and revising physical theories requires a subject that directs thought, selects among candidate explanations, and exercises judgment. To declare that subject an illusion saws off the branch on which the declaration sits. Illusions presuppose a subject that misperceives. If the subject is deleted, the word 'illusion' loses reference. The sentence 'free will is an illusion' requires a subject that can contrast seeming with being. That requirement reinstates free will.

Superdeterminism attempts to dissolve the conflict by denying the independence of measurement choices. In this view, the experimenter's decision to measure spin-up versus Bell's theorem rejects this assump-spin-down correlates with the particle's prior state through a common past. Bell's theorem assumes measurement settings can be freely (Bell, 1964) chosen. Superdeterminism rejects this assumption by claiming that every choice traces back to initial conditions that also determined the particle's properties. The

loophole saves the physics by deleting the physicist. It preserves the deterministic model by denying the very capacity—experimental choice—required to validate the model. Yet the superdeterminist still chooses which papers to write, which theories to propose, which objections to raise. Experiencing the act of advocating superdeterminism exercises the agency that superdeterminism denies.

Neuroscience experiments probe the timing of conscious will. Benjamin Libet (1985) measured electrical readiness potentials (RP) beginning 550 milliseconds before subjects reported awareness of their intention to move. The brain initiates action before conscious decision registers. Subsequent experiments refined this: Schurger (2012) showed that RP reflects general motor preparation rather than specific decision; Fried (2011) recorded individual neurons firing up to 1.5 seconds before reported awareness.

These findings constrain but do not eliminate agency. The readiness potential precedes awareness of specific intention, not the capacity for veto. Libet himself noted that subjects retain 'free won't'—the ability to cancel incipient actions after becoming aware of them. More fundamentally, experimental paradigms that measure spontaneous movements capture only a subset of willing. Deliberative decisions—weighing options, comparing outcomes, selecting among complex alternatives—unfold over seconds to hours, not milliseconds. The neuroscience of snap judgments does not generalise to the neuroscience of reflection.

Ultimately, these chronometric objections miss the mark. We do not need an oscilloscope to detect the conflict between free will and physics. The conflict is structural. If the universe is causally closed—whether deterministically or stochastically—there is no room for an uncaused cause. The incompatibility is logical, not empirical. Libet merely measured the delay of a mechanism we already knew had to exist if the brain is a physical object.

Consciousness in this context is the exercise of will on one's own stream of thought. Hold, release, redirect, compare, adopt, reject. Deliberate selection among candidate continuations. The stream is the ordered sequence of contents available for such selection. The subject is the locus at which selection is enacted.

We define commitment as the act of believing in a proposition. To rank which commitments prevail when they conflict, we define the revision cost, denoted $C(P)$, of a proposition P as the magnitude of the epistemic collapse that follows from assuming P is false. It is a measure of structural dependency. If P supports Q, then rejecting P destroys Q. The proposition with the maximal revision cost is not the one with the most evidence, but the one which provides the condition of possibility for evidence itself. Let's rank several commitments by revision cost.

I know the sky is blue. If tomorrow I learn it is an optical illusion—scattering, refraction, atmospheric tricks—fine. Mildly interesting. Nothing essential breaks.

I know that I live on Earth in the year 2025. If the simulation ends and someone unplugs me from "The Matrix", mind blown. Days to recover. But recovery is possible. I can still compare, infer, and correct.

I know there is gravity. If someone pulls the plug and reveals the simulation, forces redraw, mass no longer bends spacetime—I am stunned for weeks. I will need to rebuild the catalogue of causes and move on. The capacity to model persists.

I know $2 + 3 = 5$. If someone demonstrates that arithmetic itself is wrong—that I had a cognitive shortcut, and really $2 + 3 = 11$—the machinery of thought disassembles. Counting, comparison, consistency all rest on that foundation. Without it, reasoning collapses and is very difficult to reconstruct.

I know I have free will. I know I exist as the thing that directs its own thoughts. If this turns out to be false—then there is no 'I' left to register the failure. This is incompatible with the standpoint from which acceptability is judged.

The highest commitment dominates. Every statement, inference, or model presupposes a subject that can assert, doubt, compare, and revise. That presupposition is the content of the highest tier. Lower tiers describe states of affairs in the world. The highest tier secures the existence of the knower to whom the world appears. In any conflict, the knower wins. Without the knower, conflict is unintelligible.

Write the revision cost as $C(\cdot)$. Then $C(\text{appearances}) \ll C(\text{physics}) \ll C(\text{mathematics}) \ll C(\text{agency})$. The last inequality is decisive. If agency conflicts with physics, agency prevails. Agency is the condition for there being importance at all.

Neural correlates, receptor binding, thalamic gating, and network fragmentation describe *when* consciousness appears or vanishes and *how* physiology couples to report. That scope is exact and valuable. *What it is to be* the subject for whom appearance and vanishing matter lies elsewhere. 'When does awareness switch off?' asks about timing and mechanism. 'What is it to direct one's own thought?' asks about the standpoint that makes timing intelligible. Neuroscience answers the first. Philosophy addresses the second. Conflating them produces the reduction error: mistaking access conditions for the subject to whom access matters.

Research that maps brain states to behavioural outputs achieves correlation. Intervention studies that disrupt nodes and track changes achieve mechanism. Both are genuine progress. Constitution—the precondition without which correlation and mechanism cannot be stated—remains distinct. Consciousness sits at the constitutional level, beyond the reach of finer imaging or additional parameters.

Anaesthesia deletes awareness in seconds. Fatal familial insomnia prevents its deletion for months. Neither touches essence. We are trying to see the eye with which we see. We can map the optic nerve, treat the cataract, and measure the photon, but the act of seeing itself remains the prerequisite, not the object. Explanation terminates here, not because we have run out of data, but because we have reached the north pole.

To be fair, I must present the flip side of the coin. You can believe in free will—and ironically, I claim you have no other choice—but you also know it cannot exist, because physics is causally closed. A dichotomy with which we are forced to live.

Axiomatic Agency

Doxastic Formalism

Let S be a knowing subject, $\mathcal{P}$ the set of propositions, and $K_S \subseteq \mathcal{P}$ the commitment set of propositions S holds true. For $p, q \in K_S$, write $p \vdash q$ if q logically follows from p.

Define *revision cost*:

$$C(p) := |\{q \in K_S \mid p \vdash q\}|.$$

This induces partial order $(K_S, \preceq)$ where $p \preceq q \iff C(p) \leq C(q)$.

Hierarchy with Revision Costs

- p_1: 'Sky is blue'
  If false: Mildly interesting. Nothing breaks.
- p_2: 'Not in The Matrix'
  If false: Stunned. Rebuild ontology. Days to recover.
- p_3: 'Gravity exists'
  If false: Physics rebuilds. Weeks to recover.
- p_4: '$2 + 3 = 5$'
  If false: Arithmetic collapses. Reasoning disassembles.
- p_5: '$P \vee \neg P$' (excluded middle)
  If false: Logic fails. Cannot reason about contradictions.
- A: 'I direct my thought'
  If false: No subject remains to register the failure.

Strictly: $C(p_1) \ll C(p_2) \ll C(p_3) \ll C(p_4) \ll C(p_5) \ll C(A)$.

Agency as Maximal Element

Agency (A): capacity to perform operations on K_S (selecting, comparing, affirming, rejecting propositions). This is control over thought, not physical action.

To revise K_S by removing A requires performing an operation on K_S, which presupposes A. Thus revision of A is self-undermining:

$$A \vdash p \quad \forall p \in K_S \quad \Rightarrow \quad C(A) = |K_S|.$$

Agency is the maximal element in $(K_S, \preceq)$.

Philosophical Grounding

Descartes's Cogito ergo sum (1641): The act of doubting presupposes the existence of a doubter. Even radical scepticism cannot eliminate the thinking subject. This establishes the subject as the foundation of knowledge, not a conclusion derived from it.

Kant's Transcendental Apperception (1781): The unity of consciousness is not empirically observed but is the logical precondition for any structured experience. The 'I think' must accompany all representations. Without a unified subject, no comparison, judgment, or synthesis of data is possible.

Thomas Reid's First Principles (1785): Reid rejected both Cartesian doubt and Humean scepticism, arguing that consciousness, perception, and belief in the external world are immediate acts of common sense. They require no inferential justification because they constitute the conditions of intelligibility itself. His position anchors the self not in abstraction but in lived, self-evident awareness.

Hegel's Phenomenology of Spirit (1807): Hegel develops self-consciousness as a dialectical process—the subject becomes what it is through recognition and negation. Consciousness encounters itself in the world and, through that encounter, attains universality. Reflexivity here is not circular but generative.

David Chalmers's Hard Problem of Consciousness (1995): Chalmers formalises the explanatory gap—the difference between functional accounts and subjective experience. He frames reflexivity as evidence that consciousness is a fundamental property, not a computational artefact.

References:

Descartes, R. (1641). *Meditations on First Philosophy.*
Kant, I. (1781). *Critique of Pure Reason.*
Reid, T. (1785). *Essays on the Intellectual Powers of Man.*
Hegel, G.W.F. (1807). *Phenomenology of Spirit.*
Chalmers, D. (1995). *Facing Up to the Problem of Consciousness.*