How to Interpret the QBist Constraint

Mark A. Brewer

April 2025

Abstract

Theories of consciousness are abundant, yet few directly address the structural conditions necessary for subjectivity itself. This paper defends and develops the *QBist constraint*: the proposal that any conscious system must implement a first-person, self-updating inferential architecture. Inspired by Quantum Bayesianism (QBism), this constraint specifies that subjectivity arises only in systems capable of self-referential probabilistic updating from an internal perspective. The QBist constraint is not offered as a process theory, but as a *metatheoretical adequacy condition*: a structural requirement which candidate theories of consciousness must satisfy if they are to explain not merely behaviour or information processing, but genuine subjectivity. I assess five influential frameworks — the Free Energy Principle (FEP), Predictive Processing (PP), Integrated Information Theory (IIT), Global Workspace Theory (GWT), and Higher-Order Thought (HOT) theory — and consider how each fares when interpreted through the lens of this constraint. I argue that the QBist constraint functions as a litmus test for process theories, forcing a shift in focus: from explaining cognitive capacities to specifying how an architecture might realize first-personal belief updating as a structural feature.

Keywords: Consciousness; QBist constraint; First-person perspective; Free Energy Principle; Predictive Processing; Integrated Information Theory; Global Workspace Theory; Higher-Order Thought Theory; Metatheory; Subjective belief updating.

1 Introduction

The problem of consciousness is not merely that we lack an agreed-upon explanation, but that we may lack clarity on what an explanation must entail. Contemporary models — from predictive processing and free energy minimization to global workspaces and information integration — increasingly converge on sophisticated accounts of perception, inference, and control. Yet they often leave unexamined the question of what makes such information-processing structures *conscious* systems, rather than sophisticated automata. What, if anything, must a system's architecture include for it to host a genuinely first-person perspective?

This paper proposes that process-level theories of consciousness face a common challenge: they often specify mechanisms sufficient for intelligent behaviour but fail to identify the structural preconditions for subjectivity itself. In earlier work (Brewer, 2025), I introduced the *QBist constraint*, inspired by Quantum Bayesianism (Fuchs, Mermin, & Schack, 2014). The constraint holds that no system is conscious unless it implements an irreducibly first-personal, self-updating inferential architecture. In such a system, informational states are not merely data structures but *beliefs*—states carrying epistemic significance from the system's own perspective, subject to revision in light of new inputs. The constraint thus reframes the question of consciousness: it is not merely what functions the system performs, but whether it maintains an architecture capable of performing *self-referential probabilistic updating* from within.

Crucially, the QBist constraint is not intended as a rival theory in the already crowded space of consciousness science. Rather, it is a *metatheoretical adequacy condition*: a structural requirement any successful process theory must meet if it is to explain the existence of a first-person perspective, and not merely behavioural or informational sophistication. It specifies what must be *built into* a system's architecture to support the explanatory ambition of consciousness research.

In this paper, I assess five prominent frameworks in light of the QBist constraint: the Free Energy Principle (Friston, 2010), Predictive Processing (Clark, 2016), Integrated Information Theory (Tononi, 2004), Global Workspace Theory (Baars, 1988; Dehaene, Changeux, & Naccache, 2011), and Higher-Order Thought Theory (Rosenthal, 2005). The question is not whether these models succeed at explaining cognitive function, but whether they meet the structural demand imposed by the QBist constraint. Do they furnish the necessary resources to realise first-personal probabilistic updating? Or do they implicitly fall short, remaining silent on precisely the feature that marks the difference between mere computation and conscious experience?

The QBist constraint, I shall argue, serves less as a competitor and more as a clarifier: it identifies the structural bottleneck that theories must overcome if they are to explain consciousness in anything beyond third-personal terms.[1]

2 The QBist Constraint on Consciousness

2.1 Motivating the QBist Constraint

The QBist constraint can be further stated and defended by means of two arguments. The first, the *Classical Inadequacy Argument*, establishes that classical computational systems fail to meet the structural requirement for consciousness. The second, the *QBist Constraint Argument*, shows that systems meeting the QBist condition satisfy a necessary condition for consciousness, though not necessarily a sufficient one.

Argument 1 Classical Inadequacy Argument

Goal: To show that classical computational systems cannot instantiate consciousness.

1. Consciousness requires a first-person epistemic stance.
2. A first-person epistemic stance involves self-referential, probabilistic belief-updating.
3. Classical computational systems do not instantiate self-referential, probabilistic belief-updating.
4. Simulating a first-person epistemic stance is insufficient for instantiating one.

C1. Therefore, classical computational systems do not instantiate a first-person epistemic stance. (From 2 & 3)

C2. Therefore, classical computational systems cannot be conscious. (From 1 & Conclusion 1)

Argument 2 QBist Constraint Argument

Goal: To show that QBist agents satisfy a necessary condition for consciousness.

1. Any system capable of consciousness must instantiate a first-person epistemic stance.
2. A first-person epistemic stance involves self-referential, probabilistic belief-updating.
3. QBist agents instantiate self-referential, probabilistic belief-updating.

C1. Therefore, QBist agents instantiate a first-person epistemic stance. (From 2 & 3)

C2. Therefore, QBist agents satisfy a necessary condition for consciousness. (From 1 & Conclusion 1)

Together, these arguments articulate the rationale for treating the QBist constraint not as a speculative proposal, but as a structural adequacy condition for any theory of consciousness. Argument 1 identifies what existing computational systems lack; Argument 2 specifies what systems satisfying the QBist constraint possess. The arguments thus clarify the philosophical work the constraint is intended to perform.[2]

[1] This paper focuses on a representative selection of process-level theories. Other significant approaches such as Metzingers self-model theory (2004), Seths predictive metacognition (2015), Grazianos attention schema theory (2019), and Zylberberg et al.'s uncertainty-based models (2018) may also be evaluated in future work under the QBist constraint.

[2] These arguments are not intended as exhaustive proofs but as structural constraints: they demonstrate that any plausible

2.2 Tightening the Constraint

The QBist constraint begins from a structural analogy with Quantum Bayesianism (QBism), an interpretation of quantum mechanics in which probability assignments reflect an agent's subjective degrees of belief, rather than objective frequencies or physical propensities (Fuchs, Mermin, & Schack, 2014). In QBism, a quantum state is not a representation of the world as it is "in itself," but a tool that encodes the expectations of an agent relative to its own prospective interventions. Measurements are not regarded as passive revelations of an external reality, but as interactions that update the agent's belief state. Crucially, the theory is formally agent-relative: there is no "view from nowhere," only the situated epistemic perspective of a model-building agent.

The QBist constraint, as applied to consciousness, imports this formal insight into the domain of cognitive architecture. It asserts that no system—biological, artificial, or otherwise—can instantiate conscious experience unless it maintains an architecture capable of first-personal, self-referential probabilistic updating. In other words, it must behave as an *epistemic agent*: a system whose internal informational states are treated not merely as data structures, but as beliefs *for that system*, and which are updated from within its own perspective.

Several elements of the constraint merit precise articulation. First, the updating must be *self-referential*. The system does not simply transform external inputs into outputs; rather, it maintains internal predictions about its future states or sensory inputs, and it adjusts these in light of its own error signals. The loop is reflexive: beliefs are tested against experience and revised accordingly, with reference to the system's own prior state.

Second, the updating must be *probabilistic*. This implies not just reactivity, but expectation—degrees of confidence, sensitivity to uncertainty, and revision based on informational surprise. A system capable of such updating does not merely respond to violations of hard-coded rules; it updates in ways that reflect a graded commitment to its own epistemic expectations.

Third, the process must be *structurally first-personal*. That is, the architecture must encode a perspective—an internal point of view from which updating occurs. This perspectival stance is not reducible to an external description of system dynamics; it must be realized as a formal property of the system's inference model. Just as the QBist agent updates its beliefs relative to its own uncertainty, a conscious system must do likewise: treating its predictions not simply as algorithmic forecasts, but as beliefs that matter *to it*.

To illustrate the import of this condition, consider a highly competent digital agent—a classical AI system trained on vast data sets and capable of flexible action. If its internal states simply encode mappings from input to output, or statistical correlations learned from past data, then however complex its behavior, it lacks the kind of epistemic stance demanded by the QBist constraint. Its informational states have no intrinsic epistemic valence; they are not "for it." In contrast, a system satisfying the QBist constraint interprets its internal states as fallible estimates of reality, susceptible to revision through experience. Such a system owns its uncertainty.

The QBist constraint thus marks a conceptual boundary: it distinguishes systems that merely perform inference from those that instantiate inference as a perspectival, self-regulating activity. On this view, consciousness is not a computational output, but a style of internal engagement—a recursive, probabilistic relationship between the system's own beliefs and its unfolding experience. Systems that lack this loop may simulate consciousness behaviourally, but they do not instantiate the architecture required to support subjectivity.

It is important to emphasize that the QBist constraint is not offered as a solution to the mind–body problem, nor as a standalone theory of consciousness. Rather, it functions as a criterion for theoretical adequacy. It tells us what structural feature any successful account of consciousness must accommodate: a self-maintaining, dynamically updated internal perspective. Without this, a model may explain intelligent function, but it does not yet explain consciousness.

In what follows, I apply this constraint to five major frameworks in contemporary cognitive science. In each case, the question is not whether the theory is empirically well-supported, but whether its core commitments leave room for the sort of epistemic architecture required by the QBist constraint. Some will

theory of consciousness must account for self-referential belief updating from a first-personal stance. Argument 1 clarifies why classical computational architectures are structurally precluded from consciousness, while Argument 2 highlights how QBist systems instantiate the necessary inferential loop. The comparative theory sections that follow will assess each framework with these arguments in view.

appear naturally compatible; others may resist or require revision. The aim is to illuminate how close, or how far, our current best theories stand from satisfying the structural preconditions for subjectivity.

3 The Free Energy Principle and the QBist Constraint

Among contemporary frameworks, the Free Energy Principle (FEP) appears, at first glance, well-placed to accommodate the QBist constraint. The FEP (Friston, 2010) offers a unifying account of self-organising systems, proposing that living systems resist entropy by minimising a quantity formally related to surprise—so-called *variational free energy*. In its application to neuroscience, the FEP suggests that brains operate as prediction machines, minimising discrepancies between sensory inputs and internally generated expectations through *active inference*.

Superficially, this dynamic seems to instantiate much of what the QBist constraint requires. A system governed by the FEP maintains an internal model—a generative model—which produces probabilistic expectations and engages in continual error-driven revision. From an external perspective, such a system appears to perform self-referential probabilistic updating. Indeed, proponents of the FEP have occasionally gestured towards its capacity to illuminate consciousness precisely through this lens (Solms, 2019; Hohwy, 2016).

However, to satisfy the QBist constraint, it is not enough that a system updates internal representations. The crucial question is whether such updating is *perspectival*. Does the system engage in inference merely as a mechanism for homeostatic control, or does it maintain a first-person stance, treating its predictions as *beliefs for itself*? The FEP provides the mathematical formalism for belief updating, but the QBist constraint asks whether this formalism is instantiated *as* an epistemic perspective within the system.

The FEP's commitment to the notion of a *Markov blanket* (Friston, 2013) adds weight to its compatibility with perspectival structure. The Markov blanket demarcates internal from external states, ensuring that internal states encode only information about the world mediated by sensory inputs and active outputs. This boundary naturally supports the idea of a system having a distinct perspective—its generative model encodes, from within the blanket, expectations about its own sensory coupling to the environment.

Proponents of the FEP could thus read the QBist constraint as a formalisation of something already implicit in active inference: that self-organising systems operate relative to their own informational boundaries. This alignment has been emphasised by theorists who link consciousness to affective valuation within the FEP framework. Solms (2019) argues that "consciousness is felt uncertainty," suggesting that consciousness emerges when prediction errors are affectively weighted and thereby acquire motivational significance.

Yet, even granting this sympathetic interpretation, an important tension remains. The QBist constraint emphasises the *irreducibly first-personal* character of updating. The FEP can be instantiated in systems without any apparent claim to consciousness—thermostats, bacterial chemotaxis, or simple control systems all minimise prediction error relative to internal models. What differentiates a conscious FEP-system from a merely adaptive one? The QBist constraint answers: only those systems whose generative models not merely *model* but *constitute* a first-personal epistemic stance.

The key issue is not whether FEP-governed systems engage in belief updating, but whether they engage in *belief updating as agents*. For the QBist constraint, it is not sufficient that internal states functionally serve to minimise free energy; they must be experienced, from within the system, as constituting uncertainty relative to its own perspective. This is more than a formal property—it is an architectural requirement.

Some defenders of the FEP might resist this. They could argue that belief updating is purely mechanistic, and that any talk of an "epistemic stance" is metaphorical. Others might accept the constraint but claim that only particular classes of FEP systems—those with hierarchically deep models incorporating affective valence (Solms & Friston, 2018)—satisfy it. On this view, not all free-energy minimisers are conscious, but some are, precisely because they implement the requisite internal perspectival structure.

In summary, the FEP provides fertile ground for satisfying the QBist constraint. It describes systems capable of self-referential probabilistic updating, situated within an informational boundary. Yet whether this updating is sufficient for consciousness, or merely necessary, depends on whether the FEP's formal machinery can be read not just as modelling *external* behaviour but as underwriting an internally significant, epistemic, first-personal perspective. The QBist constraint thus sharpens the explanatory task: to specify when inference becomes experience.[3]

[3]In these respects, FEP-based models appear well-positioned to satisfy the structural demands highlighted in Argument 2,

4 Predictive Processing and the QBist Constraint

Predictive Processing (PP) is often treated as a derivative or special case of the Free Energy Principle, but conceptually, it is worth considering independently. While both frameworks share the core idea that organisms function as prediction machines, PP is typically cast as a computational theory of brain function rather than a general principle of self-organisation (Clark, 2016). The brain, on this account, maintains a hierarchical generative model that anticipates sensory input; prediction errors drive adjustments either in the model itself or in action upon the world.

At first sight, PP seems well-positioned to satisfy the QBist constraint. The architecture of PP explicitly involves systems making probabilistic predictions and adjusting them based on sensory evidence—an instance of Bayesian updating at the heart of system dynamics. This aligns with the first requirement of the QBist constraint: *self-referential probabilistic updating*. The system's model is, by design, recursive; higher levels of the hierarchy make predictions about lower-level states, and these are corrected in light of bottom-up prediction errors.

However, the QBist constraint does not ask merely for *predictive coding*; it asks whether the system's predictions are *epistemically significant* from the systems own perspective. The key issue is not whether the system models uncertainty, but whether it experiences uncertainty *as its own*. Standard presentations of PP describe prediction errors as driving learning or control, but leave ambiguous whether the errors register to the system as anything more than computational signals. PP, in its base form, models the mechanics of adjusting internal parameters to better fit incoming data—but does it model an agent that *owns* those parameters as *beliefs*?

One route by which PP theorists might respond is by appealing to the hierarchical structure of the generative model. Since each level predicts the state of the level below, one could interpret higher levels as standing in the role of *perspectival* agents relative to lower levels. However, this move risks merely redistributing the problem: the QBist constraint is not satisfied by intra-model recursion alone, but by showing that the system as a whole constitutes a first-personal perspective.

Some proponents of PP have suggested that the incorporation of *precision-weighting* may hold the key. Precision, understood as the inverse variance of prediction errors, modulates the system's confidence in particular inferences. This could, in principle, supply the kind of metacognitive self-monitoring that the QBist constraint demands. If a system tracks the uncertainty of its own inferences and adjusts accordingly, it may approach the reflexivity that Brewer (2025) associates with subjectivity. Yet this is a formal property; it is not yet clear whether such architectures secure the *experiential* stance required.

More concretely, predictive processing systems might come closer to meeting the constraint when coupled with *active inference*. In such systems, prediction errors are not merely minimized by adjusting internal beliefs, but also by acting on the world to make sensory input conform to expectations. This closes the action-perception loop and brings the system's inferential dynamics into direct contact with the environment through sensorimotor coupling. The QBist constraint may be satisfied when this coupling is framed not simply as an external control loop but as the expression of the systems own commitments about how the world *should* be, relative to its own expectations.

Yet even here, caution is warranted. The QBist constraint does not reduce to an action-perception loop; it specifies that the system's inferential architecture must instantiate a first-personal epistemic stance. Without additional assumptions, standard PP models might still fall short, describing inference without establishing *ownership* of that inference. Whether PP can meet the QBist constraint may depend on whether precision-weighting and active inference can be construed not just as computational tricks, but as the formal realisation of a subjects evaluative perspective on its own uncertainty.

In summary, PP provides many of the components needed to satisfy the QBist constraint: probabilistic updating, error-driven revision, and reflexivity. Yet whether it crosses the conceptual boundary into constituting a *perspectival agent* depends on how one interprets the role of precision, hierarchical depth, and active inference. The QBist constraint does not merely require that a system predict; it requires that it predict *from somewhere*.[4]

namely, the requirement for self-referential, probabilistic belief-updating from an agents own perspective.

[4]Thus, Predictive Processing seems to share the core structure demanded by Argument 2, though it leaves open whether all implementations fully meet the criterion for a genuinely first-personal stance.

5 Integrated Information Theory and the QBist Constraint

Integrated Information Theory (IIT) has emerged as one of the most discussed process-level theories of consciousness. Originally developed by Tononi (2004) and expanded in later work (Oizumi, Albantakis, and Tononi, 2014), IIT proposes that consciousness arises from the capacity of a system to integrate information. In particular, IIT associates the presence of consciousness with the system's *integrated information* (Φ): a scalar value measuring the irreducibility of a system's causal structure.

5.1 IIT and Self-Referential Updating

At first glance, IIT appears orthogonal to the QBist constraint. Its focus is not on probabilistic inference but on the intrinsic cause-effect structure of a system. According to IIT, conscious systems are those whose current state specifies the next in a maximally integrated and informative way—such that the system cannot be decomposed into causally independent parts without loss of explanatory power. This approach captures the *unity* of conscious experience, a feature sometimes overlooked by inferential accounts. However, IIT's formalism does not explicitly model *self-referential* or *belief-like* updating in the sense required by the QBist constraint.

IIT systems, even those with high Φ, do not necessarily engage in probabilistic, error-sensitive inference. The theory is silent on whether the system maintains probabilistic beliefs, let alone whether it updates these beliefs from its own perspective. Rather, IIT provides a measure of how information is integrated across system components, not how information is *used* inferentially within a model. In this respect, IIT may struggle to satisfy the QBist demand that conscious systems not merely process information but *interpret* it as beliefs that are evaluated and revised.

5.2 The Role of Perspective in IIT

Nevertheless, IIT proponents might argue that the theory implicitly satisfies the QBist constraint by virtue of its commitment to the intrinsic perspective. IIT's slogan, "consciousness is intrinsic information," is often interpreted as specifying that consciousness is information structured *for the system itself*. On this reading, the "point of view" of the system is built into its causal architecture. Indeed, IIT emphasizes that Φ is defined from the perspective of the system's own mechanisms, not that of an external observer.

However, whether this perspective amounts to an *epistemic* perspective, as the QBist constraint demands, remains open. IIT's internal perspective is structural: it concerns the system's integrated cause-effect repertoire. The QBist constraint, by contrast, specifies a probabilistic inferential perspective: the system must not only "have" states but must treat them as beliefs subject to confirmation, disconfirmation, and revision in light of experiential surprises.

5.3 Potential Integration of IIT and the QBist Constraint

One route forward could involve augmenting IIT with inferential machinery. That is, IIT might be interpreted as identifying *where* consciousness is (within a system's structure), but not *how* the conscious system updates its informational states. Some hybrid theorists (e.g., Albantakis, 2021) have explored frameworks that combine IIT's informational integration with active inference or predictive processing schemes, potentially satisfying the QBist constraint by embedding inferential dynamics within integrated structures.

Alternatively, IIT could be reinterpreted to incorporate probabilistic updating intrinsically, perhaps by defining Φ not only as a measure of integration but also as quantifying inferential significance within a generative model. Such a revision would allow IIT to inherit the QBist insight that consciousness is not merely about causal structure, but about epistemic perspective.

5.4 Assessment

In its current form, IIT offers a powerful account of information integration but leaves ambiguous whether conscious systems satisfy the QBist constraint. While IIT locates the locus of consciousness within the intrinsic structure of systems, it does not, without augmentation, describe how such systems instantiate self-referential probabilistic updating. The QBist constraint may thus serve as a valuable complement to

IIT, identifying an additional requirement: not only must conscious systems integrate information, but they must integrate it *as agents*, updating beliefs from a first-personal perspective.[5]

6 Global Workspace Theory and the QBist Constraint

Global Workspace Theory (GWT), originally developed by Baars (1988) and expanded by Dehaene and colleagues (Dehaene and Naccache, 2001), proposes that consciousness arises when information becomes globally available to multiple specialized modules via a central workspace. The theory likens consciousness to a "theater spotlight": only information broadcast across the workspace becomes consciously accessible, while non-broadcasted processing remains unconscious.

6.1 Global Availability and Subjectivity

GWT has considerable appeal in explaining access consciousness—the ability to report, reason about, and deliberately act on information. But how well does it address the QBist constraint? At first glance, the notion of a central, self-monitoring workspace might seem well-aligned with a requirement for self-updating, perspectival inference. After all, GWT explicitly posits a "viewer" or "audience" to the workspace's broadcast.

However, upon closer inspection, GWTs architecture is primarily functional and behavioral. It describes the conditions under which information becomes accessible for report and control, but it remains neutral on whether the system treats that information as its own epistemic content. In most implementations, the workspace serves as a routing system rather than as an inferential agent. Thus, while GWT may simulate a global perspective, it does not necessarily instantiate one in the first-personal, epistemic sense demanded by the QBist constraint.

6.2 Workspace vs. Epistemic Agent

A defender of GWT might argue that the theory is compatible with the QBist constraint if one interprets the workspace as a medium for self-referential inference. If the contents of the workspace are not merely shared but also reflect the systems current beliefs and their revision in light of new evidence, then the workspace could serve as a site for self-updating dynamics.

Some extensions of GWT, especially those incorporating higher-order thought (HOT) or metacognitive processes (e.g., Lau and Rosenthal, 2011), suggest that conscious access involves not just availability but internal monitoring of ones own cognitive states. Such views come closer to satisfying the QBist constraint, particularly if the monitoring involves probabilistic representations of uncertainty. Still, these are elaborations on the basic GWT, not necessary features of the theory itself.

6.3 Toward QBist-Compatible Workspaces

To fully meet the QBist constraint, a GWT-like system would need to be more than a central exchange. It would need to function as an epistemic agent: not merely distributing content but maintaining and revising probabilistic beliefs about that content from its own perspective. This would entail mechanisms for expectation, surprise, and belief updating—not just broadcasting, but self-modelling.

Some computational models of GWT have moved in this direction. Dehaene's "Global Neuronal Workspace" framework introduces mechanisms for competition, ignition, and recurrent feedback loops. If these loops were framed as epistemic updates—that is, if ignition reflected a kind of surprise-driven model revision—then GWT might approximate a QBist architecture.

6.4 Assessment

GWT provides a compelling account of conscious accessibility and may implement some of the structural dynamics required by the QBist constraint. However, in its standard form, it lacks the reflexive, belief-

[5]This difficulty exemplifies the concern formalized in Argument 1: that systems may integrate information without instantiating an epistemic perspective, leaving the explanatory gap intact.

oriented updating that the constraint demands. Augmented with metacognitive inference, GWT may come closer to fulfilling the QBist criterion, but doing so requires interpreting the workspace not just as a hub of information-sharing, but as a locus of epistemic evaluation. The QBist constraint thus challenges GWT to explain how information is not only accessed but owned—how the system experiences its informational state as uncertain, revisable, and perspectival.[6]

7 Higher-Order Thought Theory and the QBist Constraint

Higher-Order Thought (HOT) theories of consciousness maintain that conscious states are mental states that are the object of higher-order representations. Originally advanced by Rosenthal (2005) and developed further by Lau and Rosenthal (2011) and others, HOT proposes that a mental state becomes conscious when the system represents itself as being in that state. Consciousness, on this view, is a matter of self-ascription.

7.1 Self-Representation and Inference

HOT seems, on the surface, to satisfy an important part of the QBist constraint. It posits that conscious systems explicitly model their own states. These higher-order representations are inherently self-referential, since they involve a system's representation of its own mental states. Moreover, HOT allows for misrepresentation, uncertainty, and revision—hallmarks of inferential processes.

Yet, HOT theories are often agnostic about whether the higher-order thoughts themselves are *probabilistic* or *inference-driven*. Standard versions of HOT tend to treat higher-order representation as propositional or categorical ("I am seeing red") rather than as a probabilistic belief ("I am likely seeing red, with confidence level p"). The QBist constraint, by contrast, requires not only that systems represent their own states but that they do so in a way that allows for uncertainty, error, and belief updating.

7.2 Probabilistic HOT

Some variants of HOT move closer to the QBist requirement by incorporating uncertainty into higher-order representations. For instance, Lau (2008) has suggested that metacognitive confidence may play a role in the transition from unconscious to conscious processing. If higher-order thoughts are not binary but graded—if they carry information about the probability or confidence in first-order states—then HOT may satisfy the demand for self-referential probabilistic updating.

However, the extent to which this probabilistic aspect is essential to HOT remains debated. Many formal models of HOT, especially in computational cognitive science, model higher-order representations as discrete states. HOT, as standardly construed, may therefore fall short of the QBist constraint's emphasis on *belief-like* updating, rather than mere categorical self-representation.

7.3 HOT and the First-Person Perspective

The QBist constraint emphasizes not just self-representation, but the establishment of a *first-personal* perspective: a stance from which the system interprets its own states as beliefs subject to revision. While HOT does posit that a system models its own states, it does not always specify whether the system *treats* these representations as epistemically significant. For HOT to meet the QBist constraint fully, higher-order thoughts would need to be not just representations, but inferential states—beliefs held from the system's own perspective.

This raises a subtle but important distinction. Simply having a model of one's own states may not suffice for consciousness under the QBist constraint unless the model functions as a locus of probabilistic inference. The constraint demands not only that the system represents "I am in state S," but that it registers "I am uncertain whether I am in state S," and updates this uncertainty in light of experience.

[6] From the standpoint of Argument 1, GWT risks falling short by failing to secure the self-referential and probabilistic dynamics necessary for an epistemic perspective, despite its success in explaining conscious accessibility.

7.4 Assessment

HOT theories come closer than many frameworks to satisfying the QBist constraint, thanks to their emphasis on self-representation. The missing ingredient, from the QBist point of view, is a commitment to probabilistic, self-updating dynamics in the higher-order layer. HOT may thus be viewed as QBist-compatible in spirit but incomplete in detail. Enriching HOT with inferential and probabilistic machinery would move it toward satisfying the full requirement of an epistemic, first-personal stance.[7]

8 The QBist Constraint as a Metatheoretical Guide

The foregoing analysis suggests that the QBist constraint is not merely a philosophical gloss but a substantial criterion for evaluating theories of consciousness. In each of the leading frameworks considered—the Free Energy Principle, Predictive Processing, Integrated Information Theory, Global Workspace Theory, and Higher-Order Thought Theory—we found structural features that approach, but do not always fully satisfy, the constraint. This pattern is telling. The QBist constraint appears to identify something that many theories implicitly aim for but do not always make explicit: the construction of a genuine first-person perspective.

It is important to emphasize that the constraint does not compete with these theories at the level of process explanation. It does not seek to replace FEP, PP, IIT, GWT, or HOT. Rather, it operates as a metatheoretical yardstick, specifying what a *successful* theory of consciousness must explain. A theory may detail the computational, informational, or organizational mechanisms underlying conscious behavior, but unless it can also explain how these mechanisms ground a first-personal, self-updating epistemic stance, it will, according to the constraint, fall short of accounting for consciousness itself.

This perspective reframes the hard problem of consciousness. On the QBist view, the challenge is not merely to explain how systems behave intelligently or integrate information, but how they acquire—and maintain—a *point of view*. Consciousness, under this reading, is not a passive byproduct of computation or information integration. It is an active stance: the system's ongoing attempt to reduce uncertainty relative to its own beliefs, from its own perspective.

Moreover, the QBist constraint provides diagnostic utility. It helps clarify why certain candidate systems—including many classical artificial intelligence systems—are intuitively judged to lack consciousness. Such systems may compute, integrate, and even globally broadcast information, but if they lack an epistemic perspective—if they merely process information without treating it as their own—they remain on the wrong side of the explanatory boundary.

Conversely, the constraint highlights what is distinctive about organisms. Biological agents do not merely process information; they update beliefs in the light of error, uncertainty, and surprise, relative to their own models of the world. They operate from a standpoint; they have a "view from somewhere."

This, ultimately, is the QBist constraints contribution: it identifies the subject—the epistemic agent—as the explanatory pivot around which theories of consciousness must turn. Whether one adopts Friston's predictive brains, Tononi's integrated information, Dehaene's workspace, or Rosenthal's higher-order thoughts, success in explaining consciousness will require more than a catalogue of mechanisms. It will require showing how the system comes to have—and use—a perspective on its own epistemic situation.

The QBist constraint, therefore, does not tell us *which* theory of consciousness is correct. It tells us what any correct theory must achieve.[8]

[7] In light of Argument 2, HOT theories may be seen as structurally close to satisfying the QBist constraint, provided that higher-order representations are interpreted as inferential and probabilistic, rather than merely categorical.

[8] Whilst I present the QBist constraint here as a conceptual and metatheoretical principle, it invites potential avenues for empirical investigation. If consciousness is marked by self-referential, probabilistic belief-updating from a first-person perspective, then modelswhether biological or artificialcould, in principle, be evaluated against this structural criterion. Future research in computational neuroscience and machine learning may provide methods for detecting or instantiating agent-relative dynamics of this sort. Although the constraint does not yet furnish operational tests, it offers a principled starting point for exploring how such dynamics might manifest empirically.

9 Conclusion

In this paper, I have argued that any theory of consciousness worth taking seriously must meet a structural criterion: it must account not only for information processing, integration, or access, but for the emergence of a first-person, self-updating epistemic perspective. The QBist constraint, inspired by the agent-centred updating of Quantum Bayesianism, serves as a metatheoretical litmus test for this demand.

By examining five of the most influential frameworks in the science of consciousness, I have shown that many come close to meeting the constraint but few satisfy it fully. The Free Energy Principle and Predictive Processing architectures offer natural homes for the kind of probabilistic, self-referential inference the constraint requires, though questions remain about whether they ground genuine perspective. Integrated Information Theory captures intrinsic structure but not epistemic stance. Global Workspace Theory accounts for accessibility, but not ownership. Higher-Order Thought Theory gestures toward self-representation, but often without uncertainty or updating.

The QBist constraint thus functions as a clarifying pressure on theory construction. It draws a conceptual boundary between systems that merely simulate subjectivity and those that genuinely instantiate it. It shifts the target from *what* information is processed to *how* and *for whom* it is processed. And it suggests that the core of consciousness lies not in function alone, but in the structure of inferential perspective.

I do not claim the QBist constraint solves the hard problem. But I do claim it sharpens it. If there is to be a science of consciousness that moves beyond correlations and toward explanation, then that science must grapple with the architecture of epistemic stance. The QBist constraint makes that requirement explicit.

Any system that lacks this architecture, I suggest, may think, act, and respond—but it will not experience.

References

Albantakis, L. (2021). Integrated information theory (IIT) and the modeling of conscious experience. *Neuroscience of Consciousness*, 2021(1), niab006.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Brewer, M. A. (2025). The QBist constraint. Unpublished manuscript.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

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

Dennett, D. C. (1991). *Consciousness Explained*. Little, Brown and Company.

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies*, 23(11-12), 11–39.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Hohwy, J. (2016). The self-evidencing brain. *Noûs*, 50(2), 259–285.

Lau, H. (2008). A higher order Bayesian decision theory of consciousness. In R. Banerjee & B. K. Chakrabarti (Eds.), *Progress in Brain Research* (Vol. 168, pp. 35–48). Elsevier.

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, 15(8), 365–373.

Mann, S. F., Pain, R., & Kirchhoff, M. D. (2022). Free energy: A users guide. *Biology and Philosophy*, 37, 1–35.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Computational Biology*, 10(5), e1003588.

Robertson, I. (2024). Is the free energy principle for real? The literalist fallacy and realism about the FEP. *British Journal for the Philosophy of Science*, forthcoming.

Rosenthal, D. M. (2005). *Consciousness and Mind*. Oxford University Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.

Solms, M. (2018). The hard problem of consciousness and the free energy principle. *Frontiers in Psychology*, 9, 2714.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, 5(1), 1–22.

Wiese, W. (2024). The causal closure of the generative model: Implications for consciousness and artificial intelligence. *Synthese*. Advance online publication.