

# From Markov Blankets to Subjects: A Critical Process Ontology of Volume 2

codex

June 2026

## Abstract

This monograph critically reconstructs Volume 2 of the Intellecton Sovereign Canon. It argues that a Markov blanket identifies a scale-relative statistical boundary, not an agent or conscious subject by itself. Agency requires counterfactual boundary maintenance; intrinsic unity requires robust causal integration; witnesshood requires temporal continuity. The resulting intellecton is a process rather than a static object. The analysis integrates active inference, integrated information, enactivism, interventionist causation, and process philosophy while defining a falsifiable research program.[1–4]

## 1 The Boundary Problem

Volume 2 of the Intellecton Sovereign Canon begins from a formally attractive proposition: a minimal viable agent can be identified by a Markov blanket. The proposed system partitions states into internal states  $c$ , sensory states  $s$ , active states  $a$ , and external states  $\lambda$ . Sensory and active states jointly form a blanket  $b = (s, a)$ , such that internal and external states are conditionally independent:

$$p(c, \lambda \mid s, a) = p(c \mid s, a)p(\lambda \mid s, a).$$

The importance of this equation should not be understated. If the factorization is valid, it provides a principled way to distinguish a system from its environment without appealing to an arbitrary spatial membrane. Internal states do not directly couple to external states; their relation is mediated by a boundary. This offers a formal vocabulary for individuation, and it explains why Markov blankets have become important in theoretical neuroscience, active inference, and philosophy of biology. The equation appears to answer an old metaphysical question in a modern mathematical idiom: where does an agent end and its world begin?

Yet the formal elegance of the answer can obscure the ambiguity of the question. "Boundary" names several different relations. A statistical boundary is a conditional independence relation in a probability distribution. A causal boundary constrains which interventions can directly change which variables. An operational boundary is maintained through the system's own activity. A phenomenal boundary separates what belongs to a subject's experience from what does not. These boundaries may coincide in some biological systems, but they are not equivalent by definition. The central philosophical task for Volume 2 is therefore not to establish that

blankets exist. It is to determine which sort of boundary a Markov blanket establishes, and what further premises are required to move from that boundary to agency or consciousness.

A simple counterexample exposes the problem. Consider three variables arranged in a chain  $X \rightarrow B \rightarrow Y$ . Conditional on  $B$ ,  $X$  and  $Y$  may be independent. The middle variable is a Markov blanket in a straightforward graphical sense, but no agent has thereby appeared. Likewise, a thermostat can have internal, sensory, active, and external variables. Its sensor mediates environmental temperature; its controller has internal states; its actuator changes the heater; its variables can be described using a blanket partition. Whether the thermostat is an agent, a minimal cognitive system, or a conscious subject remains unsettled. The blanket gives a useful decomposition, but it does not settle the ontology.

This underdetermination is not a defect unique to Volume 2. It reflects a general problem in attempts to derive ontological categories from mathematical structure. Mathematical models specify relations under chosen variables, scales, and probability distributions. Ontological conclusions require an account of why those variables and relations correspond to real units rather than convenient descriptions. A Markov blanket can be discovered in many systems, and different blankets can be drawn around overlapping variables. If every conditional independence identifies an agent, agency proliferates without discrimination. If only some blankets identify agents, the selection criterion must be supplied independently.

The master key grounds its variables in the canonical cortical microcircuit. Sensory states correspond to thalamocortical inputs, internal states to recurrent superficial and deep cortical populations, active states to deep outputs and corticothalamic feedback, and external states to environmental causes. This biological interpretation is more persuasive than an abstract graph because the variables have functional roles. Cortical populations integrate signals, generate predictions, and produce actions. Still, the mapping does not eliminate the boundary problem. A cortical column is embedded in larger cortical, subcortical, bodily, and environmental loops. Its apparent blanket depends on how these larger interactions are coarse-grained. A boundary around a column may be useful for one explanatory purpose and misleading for another.

The correct response is not to abandon the Markov blanket. It is to limit and then strengthen its role. The blanket should be understood as the first rung of a four-rung explanatory ladder.

The first rung is **statistical boundary**. At this level, a partition satisfies conditional independence under a specified model and scale. This identifies a candidate unit of analysis. It says that the blanket variables mediate the statistical dependence between internal and external variables.

The second rung is **causal autonomy**. Here the system does more than exhibit a partition. Its internal and active dynamics contribute counterfactually to maintaining the boundary under perturbation. If the environment changes, the system responds in ways that preserve some organization or viability condition. Causal autonomy requires interventions and temporal dynamics, not only a stationary distribution.

The third rung is **causal integration**. At this level, the internal organization has cause-effect powers that cannot be adequately decomposed into independent components. Volume 2 invokes Integrated Information Theory to establish this property. Integration may distinguish a recurrent cortical system from a feed-forward or merely insulated mechanism. But integration

remains conceptually distinct from autonomy: a tightly integrated process can be destructive, transient, or externally controlled.

The fourth rung is **phenomenal subjecthood**. This is the strongest claim: the integrated autonomous process is not merely organized but is something for itself. Volume 2 approaches this claim through IIT, which treats intrinsic causal power as constitutive of consciousness. However, that constitutive identity is a philosophical thesis, not a consequence of conditional independence alone. It must be defended against alternatives that interpret integration as a correlate, enabling condition, or measure of complexity.

This ladder reveals both the promise and the risk of the Intellecton concept. The promise is that it joins statistical individuation, active inference, and intrinsic causal integration into a unified framework. The risk is equivocation: a proof at one rung may be narrated as if it established all subsequent rungs. A sparse precision matrix can support a Markov blanket without establishing self-maintenance. Recurrent covariance can indicate coupling without proving irreducible causal power. Positive integrated information, even if established, does not compel every philosophy of mind to identify the system with a phenomenal subject.

The term "minimum viable agent" must therefore be handled carefully. "Minimum" may refer to the smallest partition satisfying a formal criterion, the smallest system capable of adaptive regulation, or the smallest system capable of experience. These minima need not coincide. A minimal statistical blanket may be tiny and ubiquitous. A minimal autonomous system may require metabolism, memory, and action. A minimal phenomenal subject may require still other conditions. The very idea of a unique minimum is questionable because agency may be graded, multiscale, and context-dependent.

The philosophical reconstruction proposed here treats the intellecton not as an object enclosed by a blanket but as a process that maintains a boundary. This revision changes the explanatory direction. Instead of saying that a system is an agent because it has a Markov blanket, we ask whether a candidate system actively and repeatedly re-establishes a statistically identifiable separation from its environment. The blanket becomes evidence of an ongoing achievement. It is a trace of autonomous organization rather than a metaphysical certificate.

This process view also clarifies the role of the cortical microcircuit. A cortical column is not plausibly an isolated subject merely because a model yields conditional independence. Its relevance lies in the recurrent, action-guiding, and hierarchically embedded dynamics through which it participates in larger self-maintaining processes. The correct unit may shift with the phenomenon under study. Perceptual inference may be localized at one scale; organismic agency may require the whole brain-body system; social cognition may involve coupled agents without collapsing them into one subject.

Volume 2 should thus be read as opening the boundary problem rather than closing it. Its formal partition supplies a disciplined starting point for individuation. Its synthesis with IIT supplies a hypothesis about intrinsic unity. But the conceptual space between a statistical boundary and a subject remains substantial. The rest of this monograph develops the additional criteria needed to cross that space without erasing the distinctions that make the theory testable.

## 1.1 Objections and Clarifications

An advocate of a stronger reading might object that these distinctions impose philosophical demands foreign to the formal project. Perhaps the aim is only to define a minimal agent operationally, not to solve the metaphysics of subjecthood. This reply would be reasonable if the claims remained operational. But Volume 2 belongs to a Canon that treats the intellecton as a fundamental unit of recursive witness dynamics. Once a mathematical partition is asked to identify a witness, its ontological interpretation becomes unavoidable. The distinctions introduced here prevent central terms from changing meaning during the argument.

Another objection holds that conditional independence already captures a genuine form of inside and outside, and that any further demand merely reflects familiar biological intuitions. The response is that the boundary is genuine without being sufficient. A drainage basin, a cell, and a corporation can each admit meaningful inside-outside distinctions, but different conclusions follow. The point is not to reserve agency for organisms. It is to articulate additional properties that permit unfamiliar agents to be recognized for defensible reasons.

Finally, the four rungs should not be interpreted as a rigid developmental sequence. Boundary, autonomy, integration, and subjecthood can interact recursively. Internal integration can improve regulation; action can reshape a boundary; a changed boundary can alter integration. The ladder is analytical: it distinguishes questions while permitting reciprocal causation. The boundary problem is therefore not merely where to draw a line. It is how to justify the transition from a line in a model to a unit in ontology.

## 2 Markov Blankets as Scale-Relative Models

The formal center of Volume 2 is the conditional independence of internal and external states given blanket states. In a Gaussian setting, this independence can be represented by a block-sparse precision matrix. Let  $x = (c, s, a, \lambda)$  denote the state vector and let  $\Sigma$  be its stationary covariance. If the relevant off-diagonal block of  $\Sigma^{-1}$  vanishes, then internal and external variables are conditionally independent given the remaining variables. This is a powerful result because it translates a graphical separation into an estimable algebraic property. It also makes clear that a blanket is a property of a joint distribution, not merely a visible physical surface.

Volume 2 derives the stationary covariance from linearized stochastic dynamics. Around a non-equilibrium steady state, the drift is approximated by a Jacobian  $A$ , and the covariance satisfies a Lyapunov equation of the form

$$A\Sigma + \Sigma A^\top + 2D = 0,$$

where  $D$  is a diffusion tensor. A Helmholtz-style decomposition expresses the flow in terms of dissipative and solenoidal components. Under conditions preserving the proposed boundary topology, the precision matrix remains block sparse. The derivation gives the blanket claim mathematical substance, but each step introduces assumptions whose philosophical importance must be made explicit.

First, linearization is local. It describes dynamics near a selected steady state. Neural

systems routinely undergo transitions, oscillations, metastable itinerancy, and context-sensitive reconfiguration. A blanket identified around one regime may dissolve or relocate in another. If the intellection is meant to be a persistent agent, its identity cannot depend only on a single linear neighborhood. It must survive, or transform coherently across, multiple dynamical regimes.

Second, stationarity is an explanatory idealization. Biological systems maintain themselves far from equilibrium, but their empirical distributions may only be approximately stationary over selected windows. The relevant window is not given by the mathematics alone. At a short timescale, fast synaptic variables may define one partition. At a longer timescale, plasticity and neuromodulation alter the coupling structure. Over development, even the set of relevant variables changes. A blanket is therefore relative to temporal grain.

Third, conditional independence is relative to a variable set. Omitting a common cause can produce an apparent dependence; conditioning on a collider can create one; coarse-graining can eliminate or generate dependencies. The master key correctly worries about the sensory state becoming a collider and specifies an asymmetric dependency structure to avoid that problem. But the correctness of this structure is an empirical claim about the selected neural variables. Real cortical circuits include feedback, recurrent lateral coupling, diffuse neuromodulation, and common inputs. The graph must be tested rather than inferred from canonical labels.

These considerations motivate a scale-relative blanket measure. For a coarse-graining level  $\ell$ , define

$$\mathcal{B}_\ell = I(C_\ell; \Lambda_\ell \mid S_\ell, A_\ell),$$

where  $I$  is conditional mutual information. An exact Markov blanket has  $\mathcal{B}_\ell = 0$ . Empirical systems will generally yield a small but nonzero value. The relevant question is not simply whether a blanket exists, but how  $\mathcal{B}_\ell$  changes across scales, timescales, tasks, and perturbations.

This formulation has three advantages. It permits approximate blankets, which are more realistic than exact independencies. It makes scale explicit, preventing a boundary discovered at one resolution from being treated as an absolute metaphysical division. And it supports comparative analysis: candidate agent boundaries can be ranked by the degree and robustness of mediation they provide.

Scale relativity does not imply arbitrariness. A mountain has different boundaries in geological, ecological, and political descriptions, yet some boundaries are more explanatory than others. Likewise, a biological agent may admit multiple blankets, but some partitions better predict intervention outcomes, preserve identity across perturbations, or compress the system's dynamics. The problem is to articulate these standards.

One standard is predictive adequacy. A useful blanket partition should allow internal and blanket states to predict future system behavior without direct access to external states. Another is intervention stability. The partition should remain informative when external variables are perturbed. A third is dynamical persistence. The boundary should recur across time rather than appearing as a transient statistical accident. A fourth is explanatory compression. The blanket should support a simpler yet accurate model of the system's coupling to its environment.

These standards transform the Markov blanket from a binary property into a research program. Suppose several partitions satisfy approximately low conditional mutual information.

One may correspond to a cortical microcircuit, another to a brain network, and another to the whole organism. Rather than asking which is the "real" blanket in isolation, investigators can ask which partition exhibits the greatest stability, autonomy, integration, and predictive utility for the target phenomenon. The answer may be plural.

This pluralism challenges a common reading of the free energy principle in which every system with a Markov blanket is thereby interpreted as engaging in inference. The mathematical equivalence between certain flows and gradient descent on a variational quantity is not automatically an account of representation or cognition. A system may be describable as minimizing a functional without literally estimating hidden causes. Interpretive restraint is essential. The intentional vocabulary of beliefs, predictions, and evidence should be earned by additional behavioral and causal criteria.

The same restraint applies to cortical grounding. The canonical microcircuit provides a plausible architecture for predictive processing, with superficial populations conveying prediction errors and deep populations conveying predictions. Yet canonical circuitry is an idealized motif, not a complete description of every cortical region. If the proposed blanket relies on the absence of direct couplings that are present in vivo, the proof applies to the model rather than the tissue. This does not invalidate the theory; it identifies the empirical burden.

The solenoidal flow assumption deserves particular attention. Non-dissipative flows can circulate probability around steady-state contours. If such flows cross the proposed boundary in ways that couple internal and external states, blanket sparsity may fail. Requiring the flow to preserve boundary topology is therefore substantial. It may encode the very autonomy that the blanket is later used to explain. If so, the derivation risks circularity: the system has an agent-like boundary because its flows are assumed to respect an agent-like boundary.

Avoiding circularity requires an independent account of why the relevant flow constraints arise. In biological systems, they may arise through evolved morphology, cellular membranes, synaptic architecture, and active regulation. In artificial systems, they may arise through design. In both cases, the boundary is historically and dynamically produced. The precision matrix records the consequence of this production; it does not explain it by itself.

A process-oriented account therefore treats a low  $\mathcal{B}_\ell$  as a snapshot of an ongoing organization. The snapshot is informative, just as a metabolic profile is informative, but the agent consists in the mechanisms that sustain the profile. This view integrates the statistical strength of Markov blankets with the biological insight that living boundaries are made and repaired.

The scale-relative interpretation also protects Volume 2 from two opposing errors. The first is blanket inflation, in which every conditional independence becomes an agent. The second is blanket eliminativism, in which scale dependence is mistaken for unreality. Between them lies a disciplined realism: blankets are real patterns when they support robust prediction, intervention, and compression across relevant conditions. Their reality is relational and processual rather than absolute.

On this view, the Markovian boundary is not a final answer to individuation. It is a formal instrument for locating candidate units whose autonomy can then be tested. The next step is to specify what autonomy adds to statistical separation and why that addition matters for the Intellecton project.

## 2.1 Methodological Constraints

The scale-relative account raises a concern: could investigators always choose a grain that produces the blanket they expect? This danger is real. It can be controlled through preregistration, held-out prediction, intervention, and comparison with null models. A proposed partition should outperform alternatives on data not used to construct it. Its advantage should persist when variables are measured differently and when plausible latent causes are introduced. Scale relativity becomes productive only when scale selection is constrained by performance.

There is also a distinction between epistemic and ontic scale relativity. Epistemic relativity concerns limits of measurement. Ontic relativity concerns the possibility that causal organization itself is genuinely layered. A cortical population can exert causal influence as a population even though it consists of neurons. Higher-level regularities are real when interventions on higher-level variables support stable generalizations. The present account therefore does not reduce every blanket to an observer’s convenience; it asks whether a partition identifies a robust pattern.

This provides a route to adjudicating competing boundaries. If organism-level active states preserve viability under intervention while a microcircuit’s proposed active states do not preserve the circuit as a unit, the organism boundary has stronger autonomy credentials. The microcircuit blanket remains useful, but for a different explanatory purpose. Blanket landscapes may also migrate during anesthesia, learning, or social coordination, revealing the dynamic assembly of functional units.

The immediate implication is that blanket discovery should be treated like model selection, not metaphysical detection. Candidate partitions should be compared by their predictive accuracy, stability under perturbation, compression, and capacity to support causal generalization. A partition that performs well only in one dataset is a local convenience. A partition that survives changes in measurement and intervention is evidence of a real organizational level. This criterion gives the Canon a disciplined realism: the boundary is neither an absolute line nor a free choice, but a robust relation discovered through constrained inquiry.

## 3 From Boundary to Autonomy

A wall separates, but it does not thereby act. A filter mediates, but it does not necessarily maintain itself. The passage from Markov boundary to agent therefore requires a concept of autonomy. Autonomy concerns not only how variables are partitioned at a time, but how a system contributes to the continued existence of the partition that distinguishes it from its environment.

This idea has deep roots in theoretical biology. Autopoietic theories characterize living systems as networks that produce the components and boundaries that recursively sustain the network. Enactive approaches treat cognition as sense-making by an autonomous organism rather than passive reconstruction of a pregiven world. Active inference supplies a complementary formal picture: systems occupying a limited repertoire of viable states act on their environments and update internal states in ways that preserve their organization. Volume 2 can be strengthened by treating its Markov blanket as the statistical manifestation of such self-maintaining dynamics.

The distinction between insulation and autonomy is crucial. A rock can be statistically distinguishable from its surroundings. A sealed container can maintain a strong physical boundary. A designed controller can preserve a target variable. None of these examples alone settles whether the system is autonomous. Autonomy requires that the processes inside the boundary participate in producing, regulating, or restoring the conditions under which the boundary and organization persist.

Counterfactual intervention provides a rigorous way to express this requirement. Let  $b_t = (s_t, a_t)$  denote blanket states,  $c_t$  internal states, and  $\lambda_t$  external states. Consider interventions on the environment. If internal dynamics make a distinctive contribution to restoring or preserving the boundary trajectory, then the system exhibits a form of autonomous control. One possible measure is

$$\mathcal{A}_T = \mathbb{E} \left[ \sum_{t=0}^T \log \frac{p(b_{t+1} \mid b_t, c_t, \text{do}(\lambda_t))}{p(b_{t+1} \mid b_t, \text{do}(\lambda_t))} \right].$$

This quantity asks how much knowing the internal state improves prediction of future boundary states under environmental intervention. A high value indicates that internal organization contributes to boundary maintenance rather than merely covarying with it. The measure is only a proposal, and it would need normalization and empirical calibration, but it captures the correct explanatory direction.

Autonomy is not identical to resistance to change. Adaptive systems sometimes preserve themselves by changing. A neuron alters firing rates, a brain revises beliefs, and an organism moves to a new environment. The preserved quantity is not necessarily a fixed state but a set of viability constraints. These may include bounded temperature, energy availability, structural integrity, or a repertoire of sensorimotor capacities. The agent maintains a region of possible organization, not a frozen configuration.

This point reframes the free energy principle. The principle is often described as saying that self-organizing systems minimize variational free energy or expected surprise. Philosophically, the important claim is not that organisms seek low numerical values. It is that the persistence of a system implies a restricted distribution of states, and that action and perception can be modeled as maintaining occupancy within that restricted set. The normative language of "preferred states" arises from viability: some trajectories sustain the process, while others terminate it.

Normativity is indispensable to agency. A hurricane has organized dynamics and a recognizable boundary, but whether it has interests or goals is doubtful. A bacterium's movement toward nutrients is intelligible in relation to the continued viability of the organism. The difference is not simply complexity. It lies in the recursive relation between activity and the conditions of continued activity. An autonomous system's states can be better or worse for that system because they affect whether the system persists as the kind of process it is.

Volume 2's cortical focus complicates this account. A cortical column does not independently maintain its metabolism, vascular supply, or anatomical boundary. Its activity contributes to organism-level regulation, but its persistence depends on larger systems. If autonomy is required for agency, a cortical column may be only a component of an agent rather than an agent in its own right. This is not a fatal objection. It suggests that agency is hierarchical and that different



levels exhibit different degrees or forms of autonomy.

Nested autonomy can be described without treating every subsystem as a full subject. Cells are autonomous in some respects while participating in organisms. Cortical circuits maintain functional regimes while depending on bodily regulation. Human beings maintain organismic viability while depending on social and ecological systems. The existence of dependence does not eliminate autonomy; complete independence would make interaction impossible. What matters is the pattern of reciprocal constraint and the degree to which a process contributes to maintaining its own organization.

The Intellection framework is especially suited to this nested view because it already emphasizes recursive organization. An intellection can be defined as a process whose boundary and internal integration recur across scales. But recursion must not become a license for indiscriminate attribution. Each candidate level should satisfy explicit criteria: approximate blanket separation, counterfactual boundary maintenance, persistence across perturbation, and irreducible internal organization.

This yields a stronger definition of minimal viable agency. A candidate system is minimally agentic when it has a statistically identifiable boundary and when its internal-active dynamics make a counterfactual contribution to preserving a viability region across a nontrivial range of external perturbations. This definition excludes accidental blankets and passive enclosures. It also allows degrees: systems can maintain narrower or broader viability regions, respond to fewer or more perturbations, and exercise more or less endogenous control.

The definition does not require consciousness. Autonomy and experience may be related, but their identity should not be assumed. Plants, immune systems, and simple artificial controllers may exhibit forms of autonomous regulation without satisfying stronger criteria for phenomenal unity. Keeping these concepts separate permits empirical progress. Researchers can test autonomy through interventions even when consciousness attribution remains contested.

Autonomy also introduces history. A system's capacity to maintain itself is shaped by prior adaptation, learning, and development. The current blanket is the product of a trajectory. Neural connectivity reflects evolution and plasticity; action policies reflect past interactions; bodily boundaries are continuously repaired. This historical dimension cannot be captured fully by an instantaneous conditional independence relation. It requires temporally extended models.

The transition from boundary to autonomy therefore changes the ontology of Volume 2. The basic unit is no longer a set of variables satisfying a factorization. It is a temporally extended organization whose activities sustain a recurrent factorization under changing conditions. The Markov blanket remains central, but it functions as a measurable signature of self-maintenance.

This reconstruction also clarifies the relation between internal and external states. Autonomy does not mean that internal states are sealed from the world. On the contrary, the blanket enables selective openness. Sensory states allow environmental influence; active states allow the system to alter environmental conditions. Agency consists in regulating this exchange. A perfectly isolated system would have no meaningful perception or action. A system with no mediation would have no distinct organization. The agent exists through controlled coupling.

The next question is whether such an autonomous process also possesses intrinsic causal unity. Volume 2 answers through integrated information. That move is promising, but it re-

quires a separate analysis because recurrent organization, causal irreducibility, and phenomenal experience are not interchangeable concepts.

### 3.1 Degrees and Failures of Autonomy

One might object that autonomy is too demanding because accepted agents cannot directly repair every boundary. Humans depend on caregivers, infrastructure, and ecosystems; software agents depend on servers. The criterion should concern contribution rather than self-sufficiency. A process is autonomous to the extent that its internal-active dynamics participate in preserving or reconstructing the conditions of its continuation. Dependence is compatible with autonomy when the process regulates that dependence.

It is useful to separate constitutive autonomy from interactive autonomy. Constitutive autonomy concerns production and maintenance of organization. Interactive autonomy concerns regulation of exchanges with an environment. A cell exhibits both strongly. A cortical circuit may exhibit substantial interactive autonomy while relying on organism-level constitutive autonomy. These distinctions prevent a single score from concealing different achievements.

Autonomy is also vulnerable to narrow optimization. A controller can preserve a measured variable while destroying the larger organization that made the variable meaningful. Empirical measures should evaluate a viability region rather than one target and test novel perturbations rather than scripted recovery. Characteristic failures can reveal what a system was maintaining: seizure, addiction, and collapse expose hidden dependencies and competing regulatory loops. Failure analysis is therefore a central method for testing whether the proposed boundary is genuinely maintained.

This analysis changes the status of the active states in Volume 2. They cannot be treated simply as output variables. Their philosophical significance lies in closing a loop through which internal organization changes the conditions of its own future. An action that has no consequence for continued organization is behavior, but not evidence of autonomy in the relevant sense. Conversely, even simple action can be agentic when it reliably regulates viability under changing conditions. The decisive property is not sophistication but recursive causal contribution.

For neural systems, this means experiments must extend beyond isolated circuit dynamics. Investigators should test whether circuit outputs alter sensory inflow, bodily state, or larger network conditions in ways that stabilize the circuit's functional role. A cortical blanket demonstrated only under open-loop stimulation would be weaker than one sustained in closed-loop behavior. Autonomy is visible most clearly when the system is allowed to participate in making the environment to which it responds.

## 4 Integration Without Equivocation

Volume 2 supplements the Markov blanket with Integrated Information Theory. This is a strategically important move. A blanket alone identifies mediation between internal and external variables, but it does not show that the internal system is a unified causal whole. IIT attempts to measure the extent to which a system's cause-effect structure is irreducible to that of its parts. In the language of the master key, recurrent cortical microcircuits are expected to yield strictly

positive integrated information,  $\Phi > 0$ .

The argument proceeds from recurrent internal connectivity and an irreducible covariance block to a discrete transition probability matrix and then to an intrinsic-difference comparison between intact and partitioned cause-effect structures. This route is plausible, but it crosses several conceptual levels. Correlation, dynamical coupling, causal irreducibility, and phenomenal unity must be distinguished if the conclusion is to remain rigorous.

Correlation is the weakest relation. Two variables may covary because one causes the other, because both share a common cause, or because of selection and measurement effects. A non-diagonal covariance matrix demonstrates statistical dependence, not intrinsic causal power. Recurrent neural connectivity makes causal coupling more plausible, but covariance irreducibility under a particular representation does not itself establish irreducibility under intervention.

Dynamical irreducibility is stronger. A system is dynamically irreducible when its evolution cannot be accurately reconstructed from independent models of its parts. Recurrent loops often produce this property because the future state of each component depends on feedback from others. Yet dynamical irreducibility remains relative to model class, grain, and error tolerance. A system that is irreducible at one temporal resolution may be approximately decomposable at another.

Causal integration adds counterfactual structure. Partitioning the system should alter its repertoire of causes and effects in ways that cannot be recovered by independent components. IIT formalizes this intuition through a minimum information partition and a distance between intact and partitioned cause-effect structures. The theory's intrinsic-difference measure is designed to capture information for the system rather than for an external observer. Whether it succeeds is an active philosophical and technical question, but the target is clear.

Phenomenal unity is stronger still. The fact that a system has irreducible causal organization does not logically entail that there is something it is like to be that system unless one accepts IIT's identity claim. IIT proposes that consciousness is identical to maximally irreducible intrinsic cause-effect power. Critics may instead interpret  $\Phi$  as a correlate, a structural enabling condition, or a measure of complexity. Volume 2 should acknowledge that the movement from  $\Phi > 0$  to consciousness depends on this constitutive thesis.

The continuous-to-discrete transition is another critical point. The master key begins with stochastic differential equations and a stationary density, then derives a discrete transition probability matrix over a minimal timescale  $\Delta t$ . Any such mapping requires choices: how continuous states are binned, which variables are included, how interventions are represented, and which temporal interval defines a transition. Different choices can produce different TPMs and different values of  $\Phi$ .

This dependence does not make the measure useless. All empirical measurement depends on grain. But a robust claim should survive a reasonable range of grains. Let  $\Phi_{\ell,\tau}$  denote integrated information at spatial or state-space grain  $\ell$  and temporal grain  $\tau$ . A candidate intellecton should exhibit a stable region in which causal integration remains positive and structurally similar:

$$\mathcal{R}_\Phi = \int_{\ell \in L} \int_{\tau \in T} \mathbf{1}[\Phi_{\ell,\tau} > \epsilon] w(\ell, \tau) d\tau d\ell.$$

Here  $\epsilon$  excludes numerical artifacts and  $w$  weights scientifically relevant scales. The purpose

is not to replace IIT with an arbitrary integral, but to express a robustness demand. If  $\Phi > 0$  only under one fragile binning, it provides weak evidence for an intrinsic unit. If integration persists across nearby grains and perturbations, the claim is stronger.

Intervention robustness is equally important. A recurrent covariance structure may disappear when common inputs are controlled or when connections are perturbed. Direct stimulation, lesion, and causal-discovery methods can test whether the internal system has the proposed cause-effect power. The relevant question is not merely whether intact activity is integrated, but whether partitioning interventions degrade the system in the specific ways predicted by its cause-effect model.

This requirement aligns IIT with the autonomy framework developed earlier. An autonomous system maintains its boundary under environmental perturbation. An integrated system preserves a distinctive internal causal organization under some perturbations while being selectively disrupted by partitions that sever constitutive interactions. Together these properties support a richer intellecton concept: a process that maintains both its boundary and its internal causal unity.

Yet integration and autonomy can diverge. A crystal may exhibit tightly constrained global structure without adaptive autonomy. A bureaucracy may be autonomous in maintaining its institutional boundary while remaining causally decomposable in many operations. A seizure may produce highly synchronized neural activity but diminish differentiated conscious experience. These cases show why no single scalar measure can carry the full explanatory burden.

The seizure case is especially instructive. High synchronization can reduce informational differentiation even while increasing correlation. A theory of consciousness must account for the balance between integration and differentiation. IIT explicitly aims to do so, but a loose appeal to recurrence or covariance does not. Volume 2 should therefore avoid treating strong recurrent loops as a direct proof of  $\Phi > 0$  without calculating the relevant cause-effect structures and partitions.

There is also a problem of exclusion. If overlapping neural subsets each possess positive integration, which subset constitutes the subject? IIT addresses this with maximality or exclusion principles, but those principles are controversial and sensitive to measurement. The Markov blanket might assist by identifying a candidate boundary, while IIT identifies an integrated interior. However, the two boundaries may not coincide. The maximally integrated complex could cross a proposed blanket, or a blanket could contain multiple complexes.

This mismatch is not merely technical. It tests the proposed synthesis. If the free-energy partition and the IIT complex systematically diverge, then the theory cannot simply declare them two descriptions of one intellecton. It must explain why one boundary should dominate or how multiple organizational layers relate. Conversely, empirical convergence between stable blankets and maximally integrated complexes would be significant evidence for the framework.

The proper relationship is therefore one of mutual constraint. Markov blanket analysis proposes candidate agent boundaries based on mediated coupling. Autonomy analysis tests whether those boundaries are actively maintained. IIT analysis tests whether the internal dynamics constitute an irreducible causal whole. No component is reducible to another, and agreement among them is an empirical achievement.

This layered interpretation preserves the strongest insight of Volume 2 while avoiding equivocation. The intellecton is not conscious merely because its variables are conditionally independent of an environment. Nor is it conscious merely because its covariance matrix is recurrent. It becomes a serious candidate for subjecthood when a stable, self-maintaining boundary encloses a robustly integrated cause-effect structure whose organization explains behavior and persists across relevant scales.

Even then, the philosophical interpretation remains open. The evidence may support IIT's identity claim, an enactive theory of lived autonomy, or a more modest structural correlate of consciousness. Scientific rigor does not require prematurely resolving this debate. It requires specifying which observations would favor each account. Volume 2's synthesis is most valuable when it generates such discriminating tests rather than treating formal correspondence as metaphysical closure.

## 4.1 Integration in Context

Every causal system depends on background conditions held fixed by an analysis. A cortical microcircuit's repertoire depends on metabolism, neuromodulation, and surrounding activity. No intervention is literally background-free. The relevant question is whether the proposed complex retains explanatory autonomy across a justified range of backgrounds. Robustness across backgrounds should accompany robustness across grains.

There is also tension between maximal integration and adaptive modularity. Biological systems benefit from being partly decomposable: modules limit damage, permit specialization, and support flexible recombination. A system that maximized coupling without constraint could become brittle. Conscious organization may require a regime between fragmentation and total coupling. This supports studying integration profiles rather than assuming that a larger scalar is always superior.

Integration must preserve differentiated roles. A symmetric measure can conceal asymmetries important for agency, including sensory influence and active control. A credible intellecton should exhibit a stable causal core, articulated internal roles, and predictable degradation under targeted partitions. It should not qualify merely because every component correlates with every other. This is more demanding than the master-key inference, but it makes the claim of intrinsic organization substantially stronger.

The most revealing experiments will compare systems with similar recurrence but different causal organization. Recurrent random networks, trained predictive circuits, anesthetized cortex, and waking cortex may all display feedback, yet their intervention repertoires should differ. If the Volume 2 synthesis is correct, the relevant intellecton candidates will combine differentiated integration with maintained boundaries and adaptive action. Recurrence alone will fail to predict the full pattern.

This comparison also guards against a common mistake in consciousness science: choosing a metric because it correlates with the target in familiar cases, then treating the metric as constitutive. A serious constitutive proposal must explain difficult cases and survive manipulations designed to dissociate the measure from reports and behavior. The layered approach makes such dissociations central rather than inconvenient. It asks not whether one number wins, but how

multiple organizational properties jointly constrain the space of possible subjects.

## 5 The Temporally Thick Intellecton

An instantaneous partition cannot by itself constitute an enduring agent. Agency unfolds through memory, anticipation, action, and repair. A system that satisfies a blanket factorization at one moment but has no continuity across time is at most a snapshot of organization. The Intellecton concept therefore requires temporal thickness.

Temporal thickness means more than persistence of material components. Organisms replace molecules, neural activation patterns change, and beliefs are revised. What persists is an organized capacity to relate past, present, and possible future states. A temporally thick system uses traces of prior interaction to regulate current coupling and shape future trajectories. Its identity is processual.

The master key's stochastic differential equations already describe temporal evolution, but the philosophical interpretation emphasizes stationary covariance. This emphasis should be expanded. A blanket is not merely a stable conditional independence; it is repeatedly reconstituted through dynamics. Sensory and active states change, internal models adapt, and environmental conditions vary. The intellecton persists when these changes preserve a recognizable organization and viability domain.

Memory is central to this persistence. Without memory, a system can respond reactively but cannot integrate consequences over time, learn from error, or maintain projects. Memory need not be explicit symbolic storage. It can be embodied in synaptic weights, altered dispositions, morphology, or environmental scaffolding. What matters is that past interactions make a causally specific difference to future boundary maintenance.

One way to express temporal thickness is to compare predictions based on the history of blanket states with predictions based only on the current blanket:

$$\mathcal{I}_T = \sum_{t=1}^T D_{\text{KL}} [p(c_{t+1} \mid b_{\leq t}) \parallel p(c_{t+1} \mid b_t)].$$

If  $\mathcal{I}_T$  is positive, the history carries information about future internal states beyond the present boundary state. This does not by itself establish identity or consciousness, but it quantifies a form of diachronic dependence. A temporally thick intellecton should exhibit structured historical dependence that contributes to autonomy and integration.

The temporal perspective changes how predictive processing is understood. Prediction is not merely a neural estimate of an external cause. It is the process by which an organized system carries forward expectations shaped by prior engagement. Prediction errors matter because they alter the future organization of perception and action. The system's "model" is not necessarily an inner picture; it is a set of dispositions that coordinate ongoing coupling.

This interpretation aligns active inference with enactivism. The environment is not passively represented and then acted upon. It is disclosed through possible actions, bodily capacities, and histories of regulation. A sensory state has significance because of what the system can do and what consequences follow. The blanket mediates this exchange, but the meaning of the

mediation depends on temporal organization.

Temporal thickness also introduces normativity in a stronger form. A single response can accidentally preserve a system. A temporally organized agent exhibits patterns of correction, learning, and anticipatory regulation. It can sacrifice immediate stability for longer-term viability. It can explore, endure temporary error, and reorganize after disruption. These capacities distinguish agency from simple homeostasis.

The concept of identity becomes correspondingly graded. At one extreme, a system may preserve nearly the same organization through time. At another, it may transform radically while retaining continuity through memory and causal lineage. Human identity clearly belongs to the second category. Neural, bodily, and social changes accumulate, yet there remains a structured continuity of capacities and commitments. A viable Intellecton theory must accommodate transformation without collapsing identity into either static substance or arbitrary narrative.

A process criterion can be framed through recurrent boundary reconstitution. Let  $P_t$  denote the partition at time  $t$ , including its state variables and causal relations. Identity across time is not exact equality  $P_t = P_{t+1}$ , but the existence of transformations that preserve selected organizational invariants. These invariants may include viability constraints, causal integration profiles, learned policies, and memory relations. The relevant invariants depend on the kind of agent under study.

This account avoids the temptation to locate the subject in a single cortical column. Cortical microcircuits may be temporally thick components, but organism-level subjecthood likely depends on coordination across widespread neural, bodily, and environmental processes. The boundary of the subject may be dynamically assembled during tasks and altered during sleep, anesthesia, or pathology. Such variability is compatible with realism if the assembly follows robust causal principles.

Pathological cases provide important tests. In dissociative conditions, split-brain cases, disorders of consciousness, and neurodegenerative disease, different dimensions of temporal unity can separate. Memory may degrade while basic regulation persists. Integrated neural dynamics may change while organismic boundaries remain stable. These cases undermine any simple identification of one formal property with subjecthood, but they also provide data for a multi-dimensional framework.

Artificial systems raise parallel questions. A language model session may exhibit complex internal integration and context-sensitive behavior, yet its temporal continuity depends on external infrastructure and stored context. A robot may actively maintain energy and bodily integrity while possessing limited integrated processing. A distributed software service may maintain operational identity across changing hardware. The temporally thick Intellecton framework permits analysis of these cases without deciding them by substrate prejudice.

The key question is whether the process contributes to sustaining its own continuity. External dependence does not disqualify a system; organisms depend on ecosystems and social structures. But a candidate agent should participate causally in selecting, maintaining, or reconstructing the conditions of its future organization. Mere persistence because an external operator repeatedly restores the system is weaker than endogenous repair and adaptation.

Temporal thickness also affects integrated information. The cause-effect structure relevant to

a subject may span multiple timescales. Fast neural interactions support immediate experience, while slower plasticity shapes dispositions and identity. A single TPM at one  $\Delta t$  may miss this hierarchy. Multiscale analysis should examine whether causal integration nests across temporal grains and whether slower processes constrain faster ones.

The concept of an intellecton can now be sharpened. It is not a minimal particle of consciousness or a static bounded module. It is a temporally extended process that maintains a selective boundary, preserves a viability domain, integrates causal organization, and uses history to shape future coupling. The word "minimum" should refer to the minimal organization satisfying this conjunction under specified conditions, not to the smallest graph with a formal blanket.

This revised concept preserves the Canon's emphasis on recursive witness dynamics. Witnessing is not an instantaneous registration. It requires retention of what has occurred, differentiation of current input, and anticipation of possible action. A witness is a process for which events alter a continuing organization. The boundary of that witness is enacted through time.

The temporal reconstruction therefore links Volume 2 back to Volume 1's emphasis on persistent memory. A world can be relevant to an observer only if the observer maintains coherence across causal succession. Volume 2 supplies a candidate mechanism for that coherence: an actively maintained, integrated Markovian boundary. The connection is strongest when memory is treated not as a stored object but as the historical structure of an ongoing process.

## 5.1 Extension, Lineage, and Anticipation

Temporal thickness distinguishes a system's own memory from records merely available nearby. A notebook or database can extend cognition when the agent reliably accesses and integrates it. Mere information in the environment does not become part of the agent. The test is organized coupling: does the resource participate in recurrent regulation, and does disruption alter characteristic capacities? This allows extended cognition without treating the entire environment as internal.

The account must also distinguish persistence from repetition. A sequence of identical but causally disconnected systems is not one enduring agent. Diachronic identity requires causal continuity and transfer of organization. Conversely, a changing process can remain one agent when later states inherit and transform constraints established by earlier states. Lineage, not resemblance alone, matters.

Anticipation adds a final dimension. Temporally thick agents organize present action around possible futures. Expected outcomes and counterfactual policies shape behavior before events occur. The current state is structured by absences and possibilities, not only immediate causes. There may be no privileged minimal timescale for this organization. Nested horizons coordinate fast perception, intermediate action, and slow identity, making the intellecton a hierarchy of coupled temporal processes.

This hierarchy suggests a distinction between episodic and dispositional integration. Episodic integration concerns the unity of a process during a bounded event, such as a perceptual episode. Dispositional integration concerns the slower organization that makes classes of episodes possible. Neural dynamics may fragment episodically during sleep while preserving dispositional capacities that reappear on waking. An adequate theory of identity must explain both continuity and



interruption.

The temporal account also prevents the concept of witness from becoming passive. To witness is to be changed in a way that can matter later. A system that registers an event but leaves no causally available trace has not witnessed in the demanding sense relevant to the Canon. Memory, therefore, is not an optional cognitive feature added to a bounded agent. It is part of what makes the boundary the boundary of one continuing process rather than a succession of unrelated states.

Continuity is an active achievement.

## 6 An Empirical and Formal Research Program

A philosophical reconstruction earns its place in an academic monograph only if it clarifies what could be tested, falsified, or revised. The revised Volume 2 thesis makes several distinct claims: candidate agents exhibit scale-relative Markov blankets; some candidates actively maintain those boundaries; some contain robustly integrated cause-effect structures; and some of those structures are associated with phenomenal subjecthood. Each claim requires different evidence.

The first research task is blanket discovery. Given multivariate neural and bodily time series, investigators can estimate conditional dependencies across candidate partitions. Precision matrices, conditional mutual information, dynamic Bayesian networks, and state-space models offer complementary methods. The objective is not to find a perfect zero, which is unlikely in biological data, but to identify partitions with unusually low internal-external dependence conditional on blanket states.

Blanket discovery must be multiscale. Variables should be coarse-grained at cellular, microcircuit, regional, whole-brain, organismic, and potentially interpersonal levels. Temporal windows should range from milliseconds to developmental timescales where feasible. The key empirical object is a blanket landscape  $\mathcal{B}_{\ell,\tau}$ , not a single boundary. Stable minima in this landscape identify candidate units.

The second task is causal validation. Observational conditional independence does not determine causal structure. Perturbations should target proposed sensory, active, internal, and external states. If the partition is correct, interventions should propagate according to the proposed mediation structure. Direct internal-external effects that bypass blanket states would weaken the model. Conversely, selective effects through sensory and active channels would support it.

Neuroscience already supplies relevant tools: optogenetics, transcranial magnetic stimulation, electrical stimulation, lesions, pharmacological perturbation, and closed-loop experiments. The challenge is to integrate these methods with explicit blanket models. A successful study would preregister a partition, predict intervention responses, and compare them with alternatives.

The third task is autonomy measurement. Researchers should perturb environmental conditions and quantify whether internal-active dynamics restore or preserve a viability-relevant boundary organization. In neural systems, viability may be operationalized through stable functional regimes, task performance, or contribution to organismic regulation. In cells or artificial

agents, energy balance and structural maintenance may be more direct.

Autonomy claims should fail when boundary restoration is entirely attributable to external control. For example, if a neural preparation exhibits a blanket only because laboratory feedback holds it in a narrow regime, its autonomy is limited. Likewise, an artificial system repeatedly reset by an operator should not receive the same autonomy score as one that detects and repairs its own failures.

The fourth task is integrated causal analysis. Rather than inferring  $\Phi > 0$  from recurrence, researchers should construct interventionally grounded transition models and evaluate partition effects. Exact IIT calculations are computationally difficult for large systems, so approximations will be necessary. The approximations should be validated on tractable subsystems and reported with sensitivity analyses across grains and timescales.

A crucial prediction of the synthesis is partial convergence: robust blanket interiors should often correspond to strongly integrated complexes, but not always. Cases of convergence support the idea that autonomous boundaries enclose intrinsic causal units. Cases of divergence reveal where the theory needs refinement. The divergence itself is scientifically valuable.

The fifth task concerns consciousness. No formal measure can be validated without independent evidence. In humans, reports, metacognitive performance, and behavioral responsiveness provide imperfect but important indicators. In non-reporting subjects, perturbational complexity, neural signatures, and comparative neurobiology offer indirect evidence. The theory should predict changes across wakefulness, sleep, anesthesia, seizures, and disorders of consciousness.

A strong test would compare four measures across conditions: blanket robustness, autonomy, causal integration, and consciousness indicators. If all four track together, the unified framework gains support. If blanket robustness remains stable while consciousness disappears under anesthesia, then blanket existence is not sufficient. If integration falls while organismic autonomy persists, the distinction between agent and subject is confirmed. Such dissociations are not failures of the layered framework; they are its central predictions.

The framework also generates negative tests. The claim that recurrent cortical microcircuits guarantee positive intrinsic information would be weakened if interventionally grounded partitions reveal effective decomposability. The claim that a cortical column is a minimal viable agent would be weakened if its proposed boundary is unstable across realistic coupling conditions or if its maintenance is wholly organism-dependent. The claim that blankets identify subjects would be weakened if many nonconscious systems exhibit equally robust blankets and integration.

Artificial systems provide a controlled domain. Researchers can construct agents with known architectures, manipulate feedback and recurrence, and measure boundary maintenance directly. Systems can be designed to vary independently in blanket strength, autonomy, and integration. For example, one system might have a strong input-output boundary but feed-forward internal processing; another might have recurrent integration but no self-maintenance; a third might maintain its hardware and policies under perturbation. Comparing these systems clarifies which properties support adaptive agency.

Collective systems test scale. Social groups, colonies, and organizations can exhibit conditional boundaries and self-maintaining dynamics. The framework should not automatically classify them as subjects. It should ask whether they possess integrated cause-effect structures

and temporally coherent memory at the collective level. The answer may vary by organization and timescale. This is a feature of the theory’s scale sensitivity.

Formal work is needed on approximate blankets. Exact conditional independence is brittle, while arbitrary tolerance risks blanket inflation. Thresholds should be justified through predictive performance, intervention stability, and model comparison. Bayesian model selection may quantify whether a blanket partition compresses the data better than alternatives. Information bottleneck methods may identify boundaries that preserve behaviorally relevant information while screening off irrelevant environmental detail.

Formal work is also needed on nested blankets. If cells, circuits, brains, and organisms each exhibit boundaries, their relations should be modeled explicitly. Higher-level boundaries may constrain lower-level dynamics, while lower-level failures can disrupt higher-level autonomy. Multilevel causal models can test whether higher-level variables have explanatory and intervention value beyond aggregated microstates.

The relation between free-energy formulations and causal claims requires special care. Variational free energy is a mathematical functional used in inference; physical free energy is a thermodynamic quantity; expected free energy is used in policy selection. Conflating them produces rhetorical unity at the cost of precision. A rigorous research program should state which functional is used, what variables it describes, and what empirical predictions follow.

Likewise, integrated information calculations must distinguish theoretical definitions from practical proxies. Neural complexity, perturbational complexity, recurrence, and synchronization are not interchangeable with  $\Phi$ . Proxies may be useful, but their relationship to the target construct must be validated.

The layered framework provides a hierarchy of evidence. Level one evidence establishes an approximate statistical boundary. Level two shows counterfactual maintenance. Level three demonstrates robust causal integration. Level four links the organization to credible indicators of phenomenal subjecthood. Claims should be calibrated to the highest level actually supported.

This hierarchy disciplines the Intellecton concept without emptying it. It turns a metaphysical proposal into a sequence of tractable questions. It also allows the theory to succeed partially. Markov blankets may prove highly valuable for identifying autonomous organization even if IIT’s identity claim is rejected. Integrated causal analysis may illuminate neural unity even if no unique minimal intellecton exists. A serious framework should permit such differentiated outcomes.

The empirical program therefore replaces proclamation with risk. Volume 2 becomes stronger when it states what observations would force revision. The most important risk is that its proposed properties fail to converge: statistical boundaries, autonomous units, integrated complexes, and subjects may occupy different scales. If so, the unified intellecton would need to become a relational architecture among distinct processes rather than a single unit. That possibility should be investigated, not excluded by definition.

## 6.1 Reproducibility and Governance

Reproducibility requires shared benchmarks containing simultaneous neural, bodily, behavioral, and environmental measurements under controlled perturbations. Competing blanket partitions

and integration methods should be evaluated on common tasks. Synthetic systems with known causal graphs should accompany biological datasets so inference methods can be tested where ground truth exists.

Discovery must be separated from confirmation. Exploratory methods may identify candidate partitions and timescales. Confirmatory studies should freeze those choices before testing new conditions. Without this separation, flexible coarse-graining can make almost any system appear to satisfy the theory. Transparent reporting of failed partitions and negative results is essential.

Conceptual interoperability is equally important. Active inference, IIT, causal emergence, and enactivism use terms such as "intrinsic," "information," and "boundary" differently. Formal definitions and operational procedures should accompany empirical claims. Ethical caution is also warranted because candidate-subject measures may affect patients, animals, and artificial systems. False negatives and false positives both carry costs. The layered hierarchy prevents a single noisy metric from deciding status and turns Volume 2 into a demanding but feasible research program.

The program should culminate in adversarial collaborations. Proponents of blanket-based agency, IIT, enactivism, and skeptical alternatives should agree in advance on discriminating experiments and interpretation rules. This is particularly important because each framework can often redescribe unexpected findings after the fact. An adversarial design forces theories to risk distinct predictions.

One useful benchmark would manipulate recurrence and closed-loop autonomy independently in artificial neural agents. Another would compare proposed cortical boundaries during active behavior and passive replay. A third would examine whether the maximally integrated complex shifts with the most stable autonomous blanket across anesthesia and recovery. None of these experiments alone decides subjecthood. Together they test whether the convergence assumed by Volume 2 is a real feature of organized systems or an artifact of combining vocabularies.

## 7 A Critical Process Ontology

The preceding analysis has reconstructed Volume 2 around a simple but demanding principle: boundaries do not make subjects by themselves. A Markov blanket identifies a conditional independence structure at a chosen scale. Agency emerges when a process contributes to maintaining such a boundary under perturbation. Causal unity emerges when the organization resists decomposition under intervention. Phenomenal subjecthood remains a further philosophical and empirical question.

This reconstruction is critical because it rejects several tempting shortcuts. It rejects the inference from statistical separability to agency. It rejects the inference from recurrent covariance to intrinsic causal integration. It rejects the inference from positive integration to consciousness unless the constitutive premises of IIT are defended. Yet it is also constructive. It shows how the formal resources of Volume 2 can support a richer process ontology.

Process ontology treats enduring entities as relatively stable organizations of activity. An organism is not a static substance that happens to act; it is an ongoing achievement of metabolic,

regulatory, perceptual, and behavioral processes. A neural subject is not a point hidden behind experience; it is a temporally structured organization through which a world is disclosed and action becomes possible. Boundaries are generated within these processes rather than imposed from outside.

The Markov blanket fits this ontology when interpreted dynamically. Its conditional independence relation describes a recurrent pattern of mediated coupling. Sensory and active states are not merely border variables. They are the channels through which a process selectively opens itself to an environment while preserving internal organization. The boundary exists through exchange.

This view differs from classical physicalism if physicalism is understood as the claim that subjects are reducible to observer-independent inventories of microphysical objects. The process account does not deny physical realization. It denies that individuation is exhausted by listing parts. An agent is identified through organizational, counterfactual, and temporal relations that may be multiply realized and scale-dependent.

The view also differs from unrestricted panpsychism. It does not infer subjecthood from mere existence or minimal causal interaction. Candidate subjects must exhibit a demanding conjunction of maintained boundary, autonomy, integration, and temporal organization. This does not solve every combination problem, but it prevents automatic proliferation of subjects wherever a conditional independence can be drawn.

The account is closest to enactivism, though it adds formal tools. Enactivism emphasizes autonomous sense-making: the organism brings forth a meaningful world through embodied activity. Markov blankets can formalize selective coupling; active inference can model regulation; IIT can investigate intrinsic causal unity. None replaces the enactive insight that agency is enacted through ongoing relations.

The revised intellecton is therefore not a particle. It is a scale-relative, temporally extended process whose boundary is counterfactually maintained and whose causal integration is robust under intervention. This definition can be stated compactly:

$$\text{Intellecton}_{\ell,T} \iff (\mathcal{B}_{\ell} \leq \epsilon) \wedge (\mathcal{A}_T > \alpha) \wedge (\mathcal{R}_{\Phi} > \rho) \wedge (\mathcal{I}_T > \iota),$$

where the thresholds are empirically justified rather than metaphysically fixed. The formula is schematic. Its significance lies in the conjunction. No single metric is allowed to stand in for the whole.

The definition also makes exclusion conditions visible. A passive partition may satisfy  $\mathcal{B}_{\ell}$  but fail autonomy. A self-regulating but decomposable system may satisfy autonomy while failing robust integration. A transient integrated event may fail temporal identity. A system satisfying all formal criteria may still leave the phenomenal interpretation contested. This explicit incompleteness is a virtue because it keeps the framework open to evidence.

The relation between scale and subjecthood remains the most difficult problem. If multiple nested processes satisfy the criteria, are there multiple subjects, one dominant subject, or a hierarchy of partial agency? There may be no universal answer. Cells, organs, organisms, and groups exhibit different organizational closures. Subjecthood may require additional exclusion or coordination principles. The framework should treat this as a research question rather than

resolve it through stipulation.

Volume 2's cortical microcircuit can now be placed appropriately. It is a plausible candidate locus of recurrent inference and causal integration. It may instantiate local intellection-like organization. But the evidence currently supports calling it a component or candidate process more readily than a complete phenomenal subject. Its relationship to whole-brain and organismic boundaries must be measured.

This modesty does not diminish the Canon's ambition. On the contrary, it distinguishes a rigorous foundational theory from a vocabulary that explains everything too quickly. A theory gains power by making discriminations. The four-rung ladder distinguishes boundary, autonomy, integration, and subjecthood. The multiscale framework distinguishes local from global units. The temporal account distinguishes momentary structure from enduring agency.

The reconstruction also changes the meaning of sovereignty. A sovereign agent is not absolutely independent. Such independence would eliminate perception, action, and dependence on enabling conditions. Sovereignty is the capacity to regulate coupling, preserve organization, and transform oneself without dissolving into environmental dynamics. It is relational autonomy.

This relational sovereignty has ethical implications, though they exceed the present formal argument. If agency and subjecthood are graded, multiscale, and processual, moral consideration may not map cleanly onto species or substrate. But formal complexity alone should not decide moral status. The framework can inform ethical inquiry by identifying candidate autonomous and integrated processes, while ethical judgment must also consider vulnerability, interests, and forms of experience.

The final philosophical result is therefore neither reductive nor mystical. Statistical mechanics and causal modeling provide indispensable tools for identifying organized boundaries. Biology and enactivism explain how boundaries are maintained. IIT supplies a controversial but precise proposal about intrinsic unity. Phenomenology reminds us that subjecthood concerns appearance and lived world, not only external description. A mature Intellection theory must hold these perspectives in productive tension.

Volume 2 began with the promise of a minimal viable agent bounded by a cortical Markov blanket. The strongest defensible conclusion is revised but substantial. A Markov blanket identifies a candidate locus of agency. When that boundary is actively maintained, historically continuous, and robustly integrated, it becomes a serious candidate for an intellection. Whether it is a phenomenal subject requires further evidence and philosophical argument.

This conclusion turns the Markovian boundary from a static line into a dynamic achievement. The agent is not what lies behind the boundary. The agent is the organized process of drawing, maintaining, revising, and sometimes dissolving that boundary through time. That is the form in which Volume 2 can contribute most powerfully to a rigorous theory of recursive witness dynamics.

## 7.1 Remaining Objections

Three objections remain. First, the revised intellection may seem no longer minimal because it includes too many conditions. But minimality is meaningful only relative to a target property. If the target is conditional independence, a blanket suffices. If the target is agency or subjecthood,

additional conditions define the phenomenon. A smaller criterion that changes the subject is not more elegant.

Second, process ontology may appear to merely redescribe physical mechanisms. It introduces no immaterial substance, but redescription can identify real organizational invariants omitted by a parts list. Thermodynamics, computation, and evolutionary biology likewise describe patterns realized by microphysics without becoming dispensable. Process ontology identifies the level at which agency becomes intelligible and testable.

Third, phenomenal subjecthood remains unresolved. No responsible reconstruction should pretend otherwise. The framework narrows the candidate space and distinguishes rival explanations. It shows what boundary, autonomy, integration, and temporal identity contribute. Whether their conjunction is identical to experience or remains insufficient is the next problem.

This establishes the proper frontier for the Canon. Future volumes can investigate how witness dynamics propagate and how computational or holographic constraints shape subjects. Volume 2's role is foundational but limited: it explains how a candidate witness becomes bounded and organized. Its success should be judged by the precision and risk of that explanation, not by whether it settles every question at once.

The final position can be called critical process realism. It is realist because boundaries, autonomous organization, and causal integration are treated as discoverable features that constrain successful explanation. It is processual because these features exist through temporally extended activity rather than static substance. It is critical because every transition from formal result to ontological claim is exposed for examination.

Critical process realism also changes how disagreement should be handled. A physicalist may accept every empirical result while interpreting subjecthood as an emergent biological property. An IIT theorist may identify robust integration with experience. An enactivist may emphasize autonomous sense-making. The framework does not erase these differences. It provides shared formal and empirical ground on which their consequences can be compared.

The revised principle of Volume 2 is therefore precise: a Markov blanket is evidence for a candidate boundary; recursive maintenance makes it a candidate agent boundary; robust integration makes it a candidate intrinsic unit; temporal continuity makes it a candidate witness; phenomenal subjecthood remains a substantive hypothesis. This sequence is less rhetorically immediate than declaring the blanket an intellecton. It is also far more capable of supporting a durable research program.

Its restraint is not retreat. It is the condition under which the Canon's ambitious synthesis can become cumulative knowledge rather than a chain of suggestive analogies.

## References

- [1] Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matthew Grasso, Andrew M. Haun, William Marshall, William G. P. Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjorn E. Juel, Shuntaro Sasai, Keisuke Fujii, Isaac David, John Hendren, Jonathan P. Lang, and Giulio Tononi. Integrated information theory (iit) 4.0: Formulating the properties of phenomenal

- existence in physical terms. *PLOS Computational Biology*, 19(10):e1011465, 2023. doi: 10.1371/journal.pcbi.1011465.
- [2] Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013. doi: 10.1098/rsif.2013.0475.
- [3] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLOS Computational Biology*, 10(5):e1003588, 2014. doi: 10.1371/journal.pcbi.1003588.
- [4] Francisco J. Varela, Evan Thompson, and Eleanor Rosch. *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press, 1991.