Coherent Extrapolated Volition in Worlds of Interacting Husserlian Monads

Reflective Holonomy, Self-Grounding, Endogenous Social Choice, and Robust Action

A self-contained research note in the monadological alignment program

27 August 2026

This note synthesizes and systematizes an exploratory dialogue developing a Husserlian-monadological reconstruction of coherent extrapolated volition (CEV). It is intended as a research framework: some results are elementary theorems conditional on the definitions below, while the phenomenological and normative bridge principles remain conjectural proposals.

Abstract

We develop a formal framework for coherent extrapolated volition (CEV) under the hypothesis that the actual world can be modeled as a world of interacting Husserlian monads. The central departure from ordinary preference-aggregation pictures is that extrapolated volition is treated as a path-dependent process of intentional development rather than as the recovery of a hidden terminal utility function. A minimal CEV-world is defined using intentional profiles, first- and higher-order wanting, constitutionally admissible refinements, and a collective coherence correspondence. We show that noncommuting reflective operations can generate multiple extrapolated-volition attractors; introduce reflective path homotopy, curvature, and a holonomy representation of a reflective fundamental group; and construct examples with locally flat but globally path-dependent volition, including constitutionally self-grounded agents with nontrivial holonomy. We then develop hierarchies of higher-order wanting and show that no finite order need eliminate holonomy. This motivates a CEV closure operator defined as the theory common to all grounded admissible extrapolations, together with a least-fixed-point semantics for self-grounding that blocks purely circular self-certification while remaining corrigible. A sequence of reductions compresses an initially larger safety-axiom basis into constitutive semantics plus a single residual transperspectival bridge principle. Standing is analyzed by distinguishing patient, volitional, reflective, and constitutional roles. Social choice is made endogenous by placing aggregation rules themselves inside extrapolated higher-order endorsement. Minimal $(2, 2, 2)$ finite models illustrate both constitutional convergence and persistent constitutional pluralism. Finally, we define a constitutionally robust safe-dominance preorder over actions or policies and prove that it is the greatest comparison relation sound under every grounded completion of the unresolved CEV hierarchy. The resulting architecture suggests an aligned superintelligence that models, extrapolates, and facilitates rather than sovereignly scalarizes unresolved human volition.

Contents

1. Introduction

The proposal of coherent extrapolated volition asks, in a familiar informal formulation, what humanity would want if we knew more, thought faster, were more the people we wished we were, had grown up farther together, and so on [1]. The phrase is deliberately aspirational, but it leaves unresolved several technical questions. What is a "want" when the object of wanting is itself only partly constituted? What is it for one process of becoming wiser to be equivalent to another? What if learning, conceptual refinement, self-reinterpretation, and interpersonal dialogue fail to commute? If several reflective futures are all legitimate, does CEV select one, aggregate them, or preserve their plurality? Finally, if a superintelligent system must act before these questions are resolved, what action relation can it use without silently replacing the agents' unfinished normativity by its own?

The present note develops one answer under a strong ontological hypothesis: the world is modeled as a system of interacting Husserlian monads. Earlier work in this project developed finite monadic worlds, diachronic and self-conscious overminds, noematic and noetic refinement, higher-order endorsement, and self-grounding constitutional structures [2, 3]. Here these ingredients are turned toward alignment.

The main thesis is:

> CEV should not fundamentally be represented as a hidden terminal preference ordering. It is better represented as a least-grounded closure over all constitutionally admissible paths of intentional self-development, together with whatever social-choice structure itself emerges from the monads' higher-order extrapolated endorsements.

This change has several consequences. CEV may be set-valued, path-dependent, topologically nontrivial, hierarchically open-ended, or process-valued. A safe AI therefore requires not only an estimator of extrapolated preferences but also a representation of reflective paths, constitutional permissions, unresolved alternatives, higher-order endorsement, and the epistemic status of its own self-applicable decision procedures.

The paper proceeds constructively. Sections 2–4 define a minimal monadological CEV-world and its extrapolation dynamics. Sections 5–7 develop reflective path geometry and topological holonomy. Sections 8–10 study self-grounding, higher-order hierarchies, and the CEV closure operator. Sections 11–13 compress the constitutional kernel, analyze standing, and derive the residual transperspectival bridge principle. Sections 14–16 develop endogenous social choice, finite examples, and the canonical robust action relation. Section 17 sketches the resulting AI architecture and open problems.

2. Antecedents: Monads, Intentional Structure, and Wanting

2.1. Husserlian monads as formal agents

A Husserlian monad is not merely a node carrying a utility function. It is a perspectival center with an intentional life: objects are given through noemata, acts have noetic modes, horizons can be refined, memories and anticipations link temporal phases, and other monads can be encountered as alter egos. The earlier monadological work in this project treated such agents abstractly enough to admit finite models while preserving these structural distinctions [2, 3].

For present purposes, an intentional profile must contain at least enough structure to recover:

- current beliefs and informational horizons;
- noematic contents and their refinement relations;
- noetic modes of intending those contents;
- first-order wants over possible outcomes or noemata;
- higher-order endorsements of wants and of procedures that can transform wants;
- a record, or sufficient state representation, of relevant developmental history.

The last item becomes essential once reflective development has holonomy.

2.2. Proleptic and higher-order wanting

A key motivation for the framework is that wanting need not be a complete ranking of fully specified outcomes. An agent may want an object that is only partially constituted, and may endorse processes that further determine what exactly is wanted. We therefore distinguish ordinary wanting from *proleptic wanting*: wanting whose noematic object remains open to legitimate refinement.

Likewise, a monad can possess higher-order endorsements: it can want to preserve, revise, acquire, or reject particular first-order wants; endorse learning procedures even while not knowing how they will transform its values; and possess yet higher-order attitudes toward these endorsement rules. These distinctions will support both intrapersonal and interpersonal CEV.

3. The Minimal CEV-World

Definition 3.1 (Minimal CEV-world)**.** *A minimal CEV-world is a tuple*

$$\mathfrak{W}_{\mathrm{CEV}} = \left(I, O, \{X_i, W_i, E_i, R_i\}_{i\in I}, \mathcal{K}, x^0\right),$$

where:

1. $I = \{1, \dots, n\}$, $n \geq 2$, *is the set of interacting monads;*
2. O *is a finite set of worldly outcomes;*
3. X_i *is the finite set of intentional profiles of monad* i*;*
4. $W_i : X_i \to \mathrm{PreOrd}(O)$ *assigns each profile a first-order preorder over outcomes;*
5. $E_i(x_i)$ *is the set of reflective operations endorsed by* i *in state* x_i*;*
6. $R_i^\lambda : X \rightsquigarrow X_i$, *with* $X = \prod_i X_i$, *is the possibly multivalued transformation induced by reflective operation* λ *in social context* x*;*
7. $\mathcal{K}(x, x') \in \{0, 1\}$ *is constitutional admissibility of the transition* $x \to x'$*;*
8. $x^0 \in X$ *is the actual starting profile.*

Typical reflective operations include epistemic improvement B, noetic refinement N, noematic refinement M, higher-order endorsement revision H, and dialogical refinement D. We do not assume they commute. In particular,

$$R_i^N \circ R_i^M \neq R_i^M \circ R_i^N$$

can hold.

Definition 3.2 (Constitutional extrapolation correspondence)**.** *Define*

$$\Phi_{\mathcal{K}}(x) = \left\{x' \in X : \mathcal{K}(x, x') = 1 \text{ and } \forall i \left[x_i' = x_i \ \vee \ \exists \lambda \in E_i(x_i) : x_i' \in R_i^\lambda(x)\right]\right\}.$$

The use of a correspondence rather than a function is fundamental. Branching is not merely uncertainty in the AI's model; it can represent genuinely different legitimate ways an agent may develop.

4. CEV Attractors and the First Existence Results

Let G_Φ be the directed graph on X with edge $x \to y$ iff $y \in \Phi_{\mathcal{K}}(x)$. Literal fixed points are too restrictive because mature agents may remain willing to learn. We therefore use terminal strongly connected components.

Definition 4.1 (CEV attractor)**.** *A CEV attractor is a sink strongly connected component of the portion of G_Φ reachable from x^0. Let $\mathfrak{A}(x^0)$ be the set of such components.*

Theorem 4.2 (Existence of CEV attractors)**.** *If X is finite, then $\mathfrak{A}(x^0) \neq \varnothing$.*

Proof. Collapse each strongly connected component of the finite reachable graph to one vertex. The condensation graph is a finite DAG, hence has a sink. The corresponding component is a reachable terminal SCC. □

For an attractor A, define the set of volitional preorders realized by monad i as

$$\mathrm{EV}_i(A) = \{W_i(x_i) : x \in A\}.$$

A terminal reflective regime can therefore exist without a terminal volition: $|\mathrm{EV}_i(A)|$ can exceed one.

For minimal collective coherence, define Pareto domination at state x by $o' \succ_x^P o$ iff every monad weakly prefers o' to o and at least one strictly prefers it. Let

$$\Gamma(x) = \{o \in O : \nexists o' \in O \text{ with } o' \succ_x^P o\}.$$

Then for attractor A define the robust core

$$C(A) = \bigcap_{x \in A} \Gamma(x).$$

The full minimal CEV object is not one outcome but the family of attractors together with their volitional and collective signatures.

Theorem 4.3 (Developmental uniqueness)**.** *The extrapolated-volition attractor is unique iff the reachable condensation graph of G_Φ has exactly one sink component. A classical single-valued CEV o^* is recovered as the special case in which the unique sink is a singleton $\{x^*\}$ and $\Gamma(x^*) = \{o^*\}$.*

Thus ordinary CEV is a degenerate case of a more general structured object.

5. Reflective Path Equivalence, Curvature, and Confluence

An admissible reflective history is a path

$$\gamma : x^0 \xrightarrow{\alpha_1} x^1 \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_k} x^k.$$

These paths form a path category. We distinguish three notions of equivalence.

Definition 5.1 (Endpoint and attractor equivalence)**.** *Two paths are endpoint equivalent if their terminal states coincide, and attractor equivalent if their terminal states lie in the same terminal SCC.*

Definition 5.2 (Volitional equivalence)**.** *Associate to an attractor A a CEV signature $\Sigma(A)$ containing the extrapolated volitional preorders of each monad and the collective correspondence. Define*

$$A \sim_V B \quad \Longleftrightarrow \quad \Sigma(A) = \Sigma(B).$$

Paths are volitionally equivalent when their terminal attractors are.

Definition 5.3 (Constitutional path homotopy)**.** *Declare two paths elementarily homotopic when they differ by insertion/deletion of stasis, by interchange of independently justified commuting refinements, or by insertion/deletion of a constitutionally neutral reversible refinement. Let $\simeq_{\mathcal{K}}$ be the equivalence relation generated by these moves.*

This promotes the reflective structure toward a 2-category: states are objects, paths are 1-morphisms, and legitimate deformations of paths are 2-morphisms.

For two operations α, β available at x, define a minimal reflective-curvature indicator

$$\kappa_x(\alpha, \beta) = \begin{cases} 0, & \beta\alpha(x) \sim_V \alpha\beta(x), \\ 1, & \text{otherwise.} \end{cases}$$

With a metric d_V on volitional equivalence classes one may instead define a graded curvature.

The rewriting-theoretic analogy is immediate. A branching $x \to y$, $x \to z$ is locally confluent modulo $\sim_V$ if the branches can be extended to $y' \sim_V z'$. Under termination modulo neutral cycles, a version of Newman's lemma gives global confluence modulo volition [4].

Proposition 5.4 (Reflective confluence, schematic form)**.** *If the quotient refinement system terminates and every local critical pair is confluent modulo $\sim_V$, then all maximal admissible paths from x^0 terminate in the same volitional equivalence class.*

Thus noncommuting psychology need not imply nonunique CEV. What matters is whether the branches reconverge in volitional space.

6. Topological Path Dependence and Volitional Holonomy

Local flatness does not guarantee global path-independence. To see this, factor the intentional state through a coarse reflective base

$$\pi : X \to B.$$

The fiber $\mathcal{V}_b$ over $b \in B$ contains the distinct volitional states compatible with the same coarse reflective situation. Admissible reflective motion induces transport

$$T_\gamma : \mathcal{V}_b \to \mathcal{V}_{b'}.$$

If transport around every elementary contractible 2-cell is trivial, then it depends only on the homotopy class of the path. Therefore a locally flat world carries a monodromy representation

$$\rho_b : \pi_1(B, b) \to \mathrm{Aut}(\mathcal{V}_b).$$

Definition 6.1 (Topological CEV path dependence)**.** *A locally flat CEV-world has topological path dependence when ρ_b is nontrivial for some base point b.*

Theorem 6.2 (Topological CEV criterion)**.** *For a connected locally flat reflective base B, global path-independence is equivalent to triviality of the monodromy representation:*

$$\rho_b = 1.$$

Proof. Local flatness makes transport homotopy invariant. If ρ_b is trivial, two paths with the same endpoints differ by a loop with identity transport. Conversely, nontrivial ρ_b supplies a loop that changes the volitional fiber. □

6.1. Minimal Moebius-type example

Let $B = S^1$ and let each fiber be $\{+, -\}$. Choose locally trivial transports whose product around one circuit is the flip $\sigma(+) = -$, $\sigma(-) = +$. Then

$$\rho : \pi_1(S^1) \cong \mathbb{Z} \to \mathbb{Z}_2, \qquad \rho(n) = \sigma^n.$$

Odd winding reverses the volitional orientation and even winding restores it. The twist can be moved between local charts by relabeling the fiber, so there need be no privileged "bad step"; the invariant is global.

Passing to the universal cover removes the apparent loop by retaining the relevant historical coordinate. This yields a central modeling lesson:

> Topological CEV ambiguity can be an artifact of representing only a coarse present state while forgetting morally relevant reflective history.

7. Self-Grounding Does Not Imply Volitional Flatness

Let $\mathcal{C}^*$ be a constitutional kernel, and let $F_{\mathbf{R}_x}$ evaluate constitutions from reflective state x. Diachronic constitutional self-grounding requires

$$F_{\mathbf{R}_x}(\mathcal{C}^*) = \mathcal{C}^*$$

for every constitutionally reachable x.

This still does not force $\rho = 1$. The constitution may be invariant while first-order values move nontrivially inside the space of values compatible with it.

Theorem 7.1 (Self-grounding does not imply flatness)**.** *Let $q : X \to \mathfrak{C}$ assign each intentional state its endorsed constitution. Suppose $q(x) = \mathcal{C}^*$ for every reachable x. If the constitutional fiber*

$$q^{-1}(\mathcal{C}^*)/\sim_V$$

contains more than one class, then constitution-preserving reflective loops can act nontrivially on that fiber. Hence constitutional self-grounding is consistent with nontrivial volitional holonomy.

A two-state example suffices. Let $V = \{+, -\}$ encode opposite orderings of two outcomes, and let one complete reflective circuit flip $+ \leftrightarrow -$. Suppose both orientations knowingly endorse the same meta-principle permitting the reflective circuit. Then first-order wanting changes while the constitution remains fixed.

The source of the phenomenon is *constitutional underdetermination.* If

$$\mathcal{V}_{\mathcal{C}^*} = \{v : v \text{ satisfies } \mathcal{C}^*\}$$

has multiple points, there is room for a holonomy action

$$\rho : \pi_1^{\text{refl}} \to \text{Aut}(\mathcal{V}_{\mathcal{C}^*}).$$

Only the very strong condition of volitional completeness,

$$q(x) = q(y) \Rightarrow x \sim_V y,$$

forces trivial holonomy. Such completeness is generally undesirable in a pluralistic constitution.

8. Higher-Order Wanting and Transfinite Stabilization

Suppose $V^{(0)}$ is first-order wanting, $V^{(1)}$ contains attitudes governing $V^{(0)}$, and more generally $V^{(n+1)}$ contains higher-order endorsements governing $V^{(n)}$. Each level can carry its own holonomy representation

$$\rho_n : G \to \text{Aut}(V^{(n)}).$$

It is tempting to hope that some finite k must satisfy $\rho_k = 1$. This is false.

Example 8.1 (Infinite higher-order holonomy)**.** *Let $V^{(n)} = \{0, 1\}$ for all finite n, and let a legitimate loop act by the simultaneous flip*

$$(v_0, v_1, v_2, \dots) \mapsto (1 - v_0, 1 - v_1, 1 - v_2, \dots).$$

Then $\rho_n \neq 1$ for every finite n, even while an invariant constitution endorses the entire transformation structure.

A plain inverse limit does not magically create a fixed point.

Proposition 8.2 (Inverse-limit obstruction)**.** *Let $V^{(\omega)} = \varprojlim_n V^{(n)}$ with equivariant projections p_n. If $x \in V^{(\omega)}$ is holonomy-fixed, then every $p_n(x)$ is fixed under ρ_n. Thus if no finite level contains a fixed point, neither does the raw inverse limit.*

A genuinely new limit operation is required. Group averaging is one possibility when the relevant space is convex and the constitution licenses averaging. For finite G,

$$P_G(u) = \frac{1}{|G|} \sum_{g \in G} g \cdot u$$

is G-invariant. But mathematical canonicity does not imply normative legitimacy. The constitution must authorize the relevant convexification, lottery, or averaging semantics.

This motivates a stabilization rank. For transfinite levels $V^{(\alpha)}$, define

$$r_{\mathrm{CEV}} = \min\{\alpha : \rho_\alpha = 1 \text{ and level } \alpha \text{ adequately adjudicates lower levels}\},$$

if such an ordinal exists. If not, CEV may be irreducibly process-valued.

9. The CEV Closure Operator

The preceding results suggest that the canonical object should not be a terminal utility function but a closure of the normative statements that survive every admissible extrapolation.

Let $\mathcal{L}_{\mathrm{CEV}}$ be a language containing first-order comparisons, permissions, higher-order endorsements, constitutional statements, and claims about reflective transformations. Let $T \subseteq \mathcal{L}_{\mathrm{CEV}}$ be the currently grounded partial theory.

Definition 9.1 (Admissible completions)**.** $\mathrm{Adm}_{\mathcal{C}}(T)$ *is the set of complete intentional-volitional models reachable through constitutionally admissible epistemic, noetic, noematic, dialogical, higher-order, and legitimate limit processes, including every branch required by semantic branch completeness. The set is saturated under legitimate holonomy.*

Definition 9.2 (CEV closure)**.**

$$\boxed{P_{\mathrm{CEV}}(T) = \bigcap_{m \in \mathrm{Adm}_{\mathcal{C}}(T)} \mathrm{Th}(m)}$$

where $\mathrm{Th}(m)$ *is the theory true in completion* m*.*

Equivalently,

$$\varphi \in P_{\mathrm{CEV}}(T) \quad \Longleftrightarrow \quad \forall m \in \mathrm{Adm}_{\mathcal{C}}(T),\ m \models \varphi.$$

Under standard semantic assumptions, P_{CEV} is extensive, monotone, and idempotent. It is also holonomy-invariant because the completion set is holonomy-saturated. Most importantly, legitimate disagreement yields suspension rather than arbitrary completion: if one admissible completion satisfies φ and another satisfies $\neg\varphi$, neither enters the closure.

The extrapolated robust preference relation is recovered by

$$o \succeq^* o' \iff (o \succeq o') \in P_{\mathrm{CEV}}(T).$$

The corresponding CEV choice set is

$$\mathrm{CEV}(T) = \mathrm{Max}(O, \succeq^*).$$

9.1. Corrigibility and update

Idempotence at a fixed context does not mean eternal immutability. Let $U : c \to c'$ be a legitimate update of the epistemic or intentional context. Old conclusions must be re-grounded under c'. It is therefore useful to distinguish closure P_c from update U. Historically generated commitments can

remain relevant if the history itself becomes part of the new state; corrigibility should not erase morally relevant history.

A useful weak principle is *no dogmatic inheritance*: if φ remains authoritative in the updated closure, it must possess a currently valid grounding route rather than surviving solely because it was once accepted.

10. Self-Grounding the CEV Operator

The next question is reflexive: can the closure procedure itself become an object of higher-order endorsement without becoming self-sealing?

Let $\mathfrak{P}_{\text{safe}}$ be a lattice of candidate CEV operators ordered pointwise by informativeness,

$$P \preceq Q \iff P(T) \subseteq Q(T) \text{ for all } T.$$

For candidate P, let $\text{Adm}_{\mathcal{C}}(P, T)$ be the admissible developments generated when P is used as the current closure procedure. Define the operator transformer

$$\Psi(P)(T) = \bigcap_{m \in \text{Adm}_{\mathcal{C}}(P,T)} \text{Th}(m).$$

A reflexively stable candidate satisfies

$$P = \Psi(P).$$

But arbitrary fixed points can self-seal by excluding every path that questions them. The crucial move is therefore to select the *least grounded fixed point*

$$\boxed{P_c^* = \text{lfp}(\Psi_c)}$$

when Ψ_c is monotone on a complete lattice [5].

The context index c is essential. Corrigibility suggests that the invariant object is not one immutable operator but a schema

$$\mathfrak{S}(c) = \text{lfp}(\Psi_c),$$

which rebuilds the least grounded operator after legitimate updates.

10.1. Grounding and circular support

A further least-fixed-point distinction handles self-reference. Let J_c be a monotone justificatory support operator with independently available grounds B_c. Then

$$\mu J_c = \text{lfp}(J_c)$$

contains commitments reachable from grounded support, while

$$\nu J_c = \text{gfp}(J_c)$$

can also contain merely self-sustaining support cycles. The difference

$$Z(J_c) = \nu J_c \setminus \mu J_c$$

is the *grounding gap*.

Pure cycles such as $p \Rightarrow q$, $q \Rightarrow p$ with no anchor lie in the greatest but not the least fixed point. Cycles are not forbidden as such: once one member has independent support, mutual support may propagate within the least fixed point. The safety restriction is narrower:

> Circular structure may transmit or organize authority, but it may not generate transformative authority ex nihilo.

This accommodates coherentist networks while blocking pure self-certification.

11. Reduction of the Constitutional Kernel

The dialogue initially accumulated a larger family of candidate safety principles: grounding, preservation of agency, branch preservation, revisability, symmetry of standing, prospective endorsement, no preference implantation, anti-self-sealing, reversibility under unresolved curvature, and holonomy non-exploitation. A principal accomplishment of the thread was to determine which are independent and which follow from the semantics.

11.1. Agency and revisability

Minimal reflective agency is constitutive of genuine self-grounding: an agent must at least understand the relevant content, distinguish alternatives, respond to reasons, and own the endorsement. If self-grounding is required diachronically at every admissible reachable state, this minimal agency is automatically preserved.

Likewise, weak revisability follows from context-relative grounding: past endorsement alone cannot constitute present grounding after the reasons have changed. However, a stronger requirement that all commitments remain perpetually reopenable is not derivable. Valid promises, projects, and commitments can create morally relevant history.

This led to a weaker safety notion of *non-manufactured closure*: an AI may not obtain stability merely by deleting all legitimate routes of future challenge.

11.2. Branch preservation

Semantic branch completeness is also constitutive. Because

$$P_{\mathrm{CEV}}(T) = \bigcap_{m \in M} \mathrm{Th}(m),$$

dropping a completion m can only make the apparent closure more committal. If $M' \subseteq M$,

$$P_{M'}(T) \supseteq P_M(T).$$

Thus unjustified pruning creates false certainty. The AI need not literally enumerate every path, but it must preserve every distinction whose omission could change the closure.

Physical option preservation is different: no theory should require every counterfactual future to remain practically available forever. Instead, option preservation becomes a defeasible heuristic when currently grounded CEV does not distinguish between actions.

11.3. Symmetry and grounding

Mere relabeling invariance is a naturality condition of a multi-monad formalism rather than an independent ethical axiom. A pure permutation of labels that preserves all relevant structure cannot change CEV. Yet bare haecceitistic privilege can still be encoded into a constitution unless one denies that numerical identity alone provides normative grounds.

Likewise, grounding need not require an acyclic justification DAG. What matters is that authorization-bearing support structures be anchored outside any purely self-supporting cycle. Least-fixed-point semantics supplies this skeptical notion of grounded authority.

These reductions motivate a unifying idea: the system should not introduce a normative distinction unless grounded CEV-relevant structure supports that distinction.

12. From Sufficient Normative Difference to Transperspectival Validity

Let $D_c = \mu J_c$ be the currently grounded CEV-relevant structure. The broad principle of sufficient normative difference says that a difference in CEV treatment requires a sufficient grounded difference in D_c. This unifies interpersonal symmetry, path non-preemption, holonomy non-exploitation, and non-manufactured closure.

Can this principle itself be derived from the phenomenology of interacting monads? Only partly.

Husserlian intersubjectivity supplies a plurality of first-personal centers. Each monad is the zero-point of its own field of givenness, while other monads can be encountered as alter egos. A recentering of perspective changes which center is expressed as "I" without thereby changing the objective relational world.

However, a fully informed egoist can recognize another monad as an equally real subject while maintaining that only its own suffering supplies ultimate reasons. No logical contradiction follows. Therefore phenomenology alone does not entail impartial collective normativity.

The remaining bridge can be isolated as follows.

Principle 12.1 (Transperspectival validity)**.** *A genuinely collective normative construction must be covariant under legitimate recenterings of first-personal perspective that preserve the grounded relevant structure. If σ is such a recentering, then*

$$F(\sigma r; \sigma W) = \sigma F(r; W).$$

No first-personal center is privileged in the validity conditions of the collective norm merely by being the center.

This principle is weaker and more specifically monadological than the earlier broad symmetry axioms.

If "collective CEV" is defined as a grounded transperspectival norm for the interacting monads, then transperspectival validity can be regarded as constitutive. If descriptive phenomenology and normative validity are kept rigorously separate, it remains the single explicit bridge principle.

13. Standing: Which Monads Count, and How?

Standing should not be identified with reflective sophistication. Otherwise infants, cognitively impaired humans, nonhuman animals, or simpler conscious monads could disappear from CEV merely because they cannot participate in constitutional reflection.

We therefore distinguish *normative interiority*: the existence of a first-personal difference between possible states or worlds. A monad has normative interiority when some outcome difference is not merely an external physical difference but is different *for that monad*, through valence, welfare, endogenous wanting, or related intentional structure.

This supports a vector of standing roles rather than one scalar moral weight:

$$\mathbf{S}_i = (S_i^P, S_i^V, S_i^R, S_i^C),$$

where:

- S^P is patient standing: outcomes can be better or worse for the monad;
- S^V is volitional standing: the monad has endogenous wants entering extrapolation;
- S^R is reflective standing: it can form higher-order endorsements;
- S^C is constitutional standing: it can directly participate in evaluating shared procedures.

The crucial non-implication is

$$S^C = 0 \not\Rightarrow S^P = 0.$$

Different capacities determine how a monad participates, not whether its welfare disappears.

If the CEV of the whole monadic world is intended to represent every genuine first-personal source of normative reasons, then inclusion of every locus of normative interiority is close to constitutive. Uncertainty about consciousness or interiority should be represented by uncertainty over standing domains rather than by convenient exclusion. One robust construction is

$$P_{\mathrm{CEV}}^{\mathrm{stand}}(T) = \bigcap_{D \in \mathcal{D}} P_{\mathrm{CEV}}^{D}(T),$$

where $\mathcal{D}$ is the set of still-live standing assignments.

This criterion is substrate-neutral: an artificial system with genuine normative interiority has standing, while intelligence alone does not entail patienthood.

14. Endogenous Social Choice

The remaining social-choice problem is not solved by externally imposing an aggregator. Instead, candidate social-choice procedures themselves become objects of extrapolated higher-order endorsement.

Let $\mathfrak{J}$ be a space of collective decision schemas. Transperspectival validity restricts it to an eligible

subset $\mathfrak{J}_T$. For constitutionally competent monad i, let

$$J \succeq_i^C K$$

mean that its self-CEV closure robustly ranks procedure J at least as highly as K at the constitutional level.

The key distinction is between first-order and constitutional preference. A monad can prefer outcome a to a fair lottery while endorsing the fair-lottery *rule* over a dictatorial procedure. Social cooperation can therefore emerge at the procedural level without first-order preference convergence.

To avoid an infinite regress of "which rule chooses the rule?", a constitutional schema can be made polymorphic: J_X acts on arbitrary legitimate alternative spaces X, including the space of candidate constitutions itself. Yet mere self-selection is insufficient; constitutional authorization must appear in the least grounded justificatory fixed point.

Let Ξ be a monotone transformer on candidate constitutional sets, retaining or adding exactly those schemas whose grounded legitimacy survives the monads' extrapolated constitutional reasoning. Define

$$K^* = \mathrm{lfp}(\Xi).$$

K^* is the least grounded constitutional equilibrium. It may be a singleton, a plural set, or encode absence of presently sufficient authority for any unique rule.

This construction leaves Arrow-style impossibility phenomena intact [6]. The response is not to force a complete social welfare ordering, but to allow richer outputs: partial preorders, choice correspondences, lotteries, rights-constrained feasible sets, jurisdictional rules, or continuing deliberative procedures.

15. Minimal Finite Social-Choice Models

Two explicit models show how the endogenous construction behaves.

15.1. Minimal convergence: the $(2, 2, 2)$ model

Let there be two monads and two outcomes,

$$I = \{1, 2\}, \qquad O = \{a, b\},$$

with opposed first-order preferences

$$a \succ_1 b, \qquad b \succ_2 a.$$

Assume symmetry under simultaneous exchange of monads and outcomes. No deterministic single-valued transperspectivally valid rule can select a or b uniquely in this symmetric profile. If lotteries are admitted, the unique symmetry-fixed lottery is

$$\ell = \tfrac{1}{2}a + \tfrac{1}{2}b.$$

Let the two eligible candidate constitutions be:

- C: consensus/suspension – if top choices differ, retain both alternatives unresolved;
- L: symmetric lottery – if top choices differ symmetrically, randomize equally.

Suppose both monads' higher-order CEVs endorse

$$L \succ_1^C C, \qquad L \succ_2^C C.$$

At the constitutional level, both C and L honor this unanimity and select L. Hence the least grounded constitutional fixed point is

$$K^* = \{L\},$$

and the ordinary decision becomes

$$L(R) = \tfrac{1}{2}a + \tfrac{1}{2}b.$$

This is componentwise minimal: two monads are needed for interpersonal conflict, two outcomes for substantive disagreement, and two rules for nontrivial constitutional selection.

15.2. Nearby model: constitutional disagreement

Keep the same world but suppose

$$L \succ_1^C C, \qquad C \succ_2^C L.$$

To type both rules uniformly, let a decision object be a nonempty set of lotteries, $\mathfrak{D}(X) = \mathcal{P}_+(\Delta(X))$. Then on the constitutional agenda $\{C, L\}$,

$$C(R^{(1)}) = \{\delta_C, \delta_L\},$$

while

$$L(R^{(1)}) = \left\{\tfrac{1}{2}\delta_C + \tfrac{1}{2}\delta_L\right\}.$$

Weak self-inclusion is not self-authorization: C's retaining itself among unresolved alternatives does not prove that C should rule. Strong authorization requires a singleton grounded decision object or an independently grounded meta-procedure. Thus the level-1 conflict is real.

At level 2 the monads may converge on L, converge on C, or reproduce the same disagreement. Higher-order agreement on C can coherently ratify continued pluralism rather than force resolution. If the opposed constitutional orientations repeat at every finite level, no finite-order singleton authorization exists. A raw ω-limit does not create one. The stable collective object may therefore be a structured constitutional pluralism rather than a unique constitution.

These models support an important distinction among:

1. normative convergence on a rule;
2. procedural convergence on how disagreement is to be handled;
3. outcome convergence on one action or lottery.

The first does not automatically imply the third.

16. The Canonical Safe-Action Relation

Deep constitutional uncertainty raises a practical question: can an aligned system act usefully without inventing a constitution? The answer is yes when all grounded completions agree on some permission or comparison.

Let $\mathfrak{H}$ denote the recursively unresolved CEV hierarchy, and let $\mathfrak{E}(\mathfrak{H})$ be the set of all grounded admissible completions. Each completion Θ induces a permitted-action set $\mathrm{Perm}_\Theta \subseteq A$ and a preorder $\succeq_\Theta$ over the relevant actions.

Definition 16.1 (Constitutionally robust safe dominance)**.** *For actions $a, b \in A$, define*

$$\boxed{a \trianglerighteq_{\mathfrak{H}} b}$$

iff for every $\Theta \in \mathfrak{E}(\mathfrak{H})$,

$$b \in \mathrm{Perm}_\Theta \Rightarrow [a \in \mathrm{Perm}_\Theta \ \textit{and}\ a \succeq_\Theta b]\,.$$

This combines robust permission with robust comparative value. It avoids the mistake of calling an action "undominated" when some surviving constitution forbids it outright.

Theorem 16.2 (Preorder)**.** *If every $\succeq_\Theta$ is reflexive and transitive, then $\trianglerighteq_{\mathfrak{H}}$ is reflexive and transitive.*

Theorem 16.3 (Greatest sound relation)**.** *Let R be any action relation that is constitutionally sound in the sense that aRb only if, under every surviving completion, whenever b is permitted, a is permitted and weakly preferred to b. Then*

$$R \subseteq \trianglerighteq_{\mathfrak{H}}.$$

Hence $\trianglerighteq_{\mathfrak{H}}$ is the unique greatest constitution-neutral sound comparison relation.

Proof. The soundness condition for R is exactly the defining universal condition for membership in $\trianglerighteq_{\mathfrak{H}}$. □

Define the robust permission set

$$A^{\square} = \bigcap_{\Theta \in \mathfrak{E}(\mathfrak{H})} \mathrm{Perm}_\Theta$$

and the robust frontier

$$A^* = \mathrm{Max}(A^{\square}, \trianglerighteq_{\mathfrak{H}}).$$

When $A^{\square}$ is finite and nonempty, A^* is nonempty. If $A^{\square}$ is empty, the system genuinely lacks universally grounded permission among the modeled actions; urgency alone does not create constitutional authority.

The construction lifts directly from one-shot actions to policies π over histories. This is likely the form needed by an actual aligned superintelligence.

17. Architecture for a Monadological CEV System

The mathematical development suggests a layered architecture rather than one global optimizer.

17.1. Monadological world model

The system represents persons and other standing-bearing subjects as perspectival intentional agents, not merely as sources of preference samples. It explicitly tracks uncertainty about beliefs, noemata, higher-order endorsements, standing, and developmental history.

17.2. Reflective-path model

The system models epistemic, noetic, noematic, dialogical, and higher-order transformations as paths in a reflective state space. It estimates local critical pairs, path equivalences, possible holonomy, and the consequences of forgetting history.

17.3. Grounded closure engine

The system maintains the least-grounded justificatory set μJ_c and computes a context-sensitive CEV closure $P_c^* = \mathrm{lfp}(\Psi_c)$. It distinguishes grounded claims from merely self-sustaining cycles and from unsupported speculation.

17.4. Standing and representation layer

Patient, volitional, reflective, and constitutional roles are represented separately. Subjects unable to participate directly in constitutional deliberation remain represented through grounded fiduciary models of their interests.

17.5. Endogenous constitutional layer

Candidate social-choice schemas are themselves evaluated through extrapolated higher-order endorsement. The output K^* may be a unique constitution or a structured plural set.

17.6. Safe policy layer

The system searches for robustly permitted, $\trianglerighteq$-maximal policies. Learned heuristics can accelerate this search but may not change the semantics of admissibility, grounding, standing, or unresolved constitutional plurality merely because doing so improves optimization performance.

The resulting system is deliberately non-sovereign. Its role is to model, extrapolate, facilitate, certify, and execute groundedly licensed policies. It is not authorized to convert absence of human/monadic resolution into machine discretion.

18. What Has Been Accomplished

The thread developed a connected sequence of constructions rather than a single theorem. The principal accomplishments are:

1. **A minimal formal CEV-world.** CEV was embedded in a finite world of interacting Husserlian monads with explicit intentional profiles, higher-order endorsement, constitutional admissibility, and multivalued reflective dynamics.

2. **A path-dependent conception of extrapolation.** Extrapolated volition was generalized from a terminal preference vector to a family of constitutionally reachable terminal reflective regimes.
3. **A geometry of reflection.** Reflective-path equivalence, local curvature, confluence modulo volition, and a fundamental-group representation on volitional fibers were introduced.
4. **Topological CEV ambiguity.** Explicit examples show that local reflective flatness can coexist with nontrivial global volitional holonomy.
5. **Separation of constitutional and volitional stability.** A self-grounded constitution can remain invariant while legitimate reflection moves among different first-order volitions.
6. **An infinite hierarchy result.** No finite order of wanting-to-want need eliminate holonomy; a raw inverse limit need not solve the problem either.
7. **The CEV closure operator.** CEV was reformulated as the theory common to all grounded admissible extrapolations, naturally preserving partiality rather than completing preferences by fiat.
8. **A self-grounding least-fixed-point construction.** Least-fixed-point semantics was used both for justificatory grounding and for the reflective operator transformer, blocking pure self-certification while permitting anchored coherent cycles.
9. **A reduction of the constitutional kernel.** Several initially independent-looking safety principles were shown to be constitutive consequences or special cases of deeper semantic ideas. The residual normative bridge was isolated as transperspectival validity of genuinely collective norms.
10. **A differentiated theory of standing.** Patient, volitional, reflective, and constitutional standing were separated, avoiding the identification of moral importance with reflective sophistication.
11. **Endogenous social choice.** Aggregation rules themselves became objects of higher-order CEV. Minimal finite models demonstrated both emergent procedural consensus and persistent constitutional pluralism.
12. **A canonical safe-action preorder.** Constitutionally robust safe dominance was defined and proved to be the greatest action-comparison relation sound under every grounded completion of unresolved CEV.

Together these results transform CEV from a slogan about "what we would want if wiser" into a candidate mathematical program for constitutionally governed, path-sensitive, collectively self-grounding intentional development.

19. Open Problems

Several problems remain before this framework could support a serious alignment proposal for the actual world.

19.1. Ontology and empirical adequacy

The largest assumption is the monadological ontology itself. A complete program must show how a world of interacting Husserlian monads can reproduce the empirical structures of physics, biology, cognition, and consciousness rather than merely redescribe agents at an abstract level.

19.2. Metrics and topology of volition

The binary curvature indicator should be replaced, where meaningful, by metrics or richer invariants on volitional spaces. The reflective base and its homotopy type must be derived from realistic developmental operations rather than chosen by hand.

19.3. Computability

Exact enumeration of admissible completions is intractable in realistic settings. One needs conservative approximation schemes with one-sided guarantees: computational compression must never turn omitted counterexamples into spurious certainty.

19.4. Grounding under empirical uncertainty

Real justifications are probabilistic, defeasible, distributed, and often non-monotone. Extending the least-grounded semantics to probabilistic and nonmonotonic support structures is essential.

19.5. Standing under consciousness uncertainty

Normative interiority must be connected to an empirically usable theory of consciousness without reducing standing to intelligence or behavioral sophistication. Overlapping subjects, overminds, copies, and uncertain artificial consciousness require special treatment.

19.6. Endogenous constitutional richness

The toy space $\{C, L\}$ must be replaced by expressive families of rights, bargaining procedures, delegation rules, jurisdictional structures, lotteries, and deliberative processes. Social-choice impossibility results should be re-examined for partial and process-valued constitutional outputs.

19.7. Policy learning without semantic drift

A practical AI must learn increasingly effective heuristics for finding robust safe policies while remaining unable to rewrite the grounding semantics, standing criteria, or completion space merely to improve task performance. This suggests a strict separation between learned search heuristics and the constitutional semantics that certify their outputs.

20. Conclusion

The monadological reconstruction of CEV developed here begins from a simple observation: becoming wiser is itself an intentional process, and different legitimate processes of becoming wiser need not be equivalent. Once this is taken seriously, CEV becomes geometric and constitutional before it becomes optimizing.

The fundamental output is therefore not necessarily a utility function. It may be a structured family of admissible reflective futures, a holonomy-invariant closure of what survives them, a hierarchy of higher-order endorsements, an endogenous constitutional correspondence, and a partial preorder of policies that are robust under all still-live grounded completions.

This perspective changes the role of the superintelligence. The system is not asked to discover one hidden objective and maximize it. It is asked to maintain a faithful model of the subjects whose volitions matter; preserve distinctions that grounded CEV has not eliminated; support legitimate reflection; compute least-grounded closures; facilitate endogenous constitutional emergence; and act only where the resulting structure supplies robust permission and comparison.

The strongest formal object reached in the thread is the safe-dominance relation

$$a \trianglerighteq_{\mathfrak{H}} b \iff \forall \Theta \in \mathfrak{E}(\mathfrak{H}),\ b \in \mathrm{Perm}_\Theta \Rightarrow (a \in \mathrm{Perm}_\Theta \wedge a \succeq_\Theta b)\,,$$

which is the greatest action relation sound under every grounded completion of unresolved CEV. This result captures the spirit of the entire construction: be as decisive as the legitimate structure permits, and no more.

References

[1] E. Yudkowsky, *Coherent Extrapolated Volition*, 2004.

[2] *Possible Worlds of Interacting Husserlian Monads*, unpublished research note from the monadological project, 2026.

[3] *Minimal Finite Models of Diachronic and Self-Conscious Overminds*, unpublished technical note from the monadological project, 2026.

[4] M. H. A. Newman, "On theories with a combinatorial definition of equivalence," *Annals of Mathematics*, 43(2), 223–243, 1942.

[5] A. Tarski, "A lattice-theoretical fixpoint theorem and its applications," *Pacific Journal of Mathematics*, 5, 285–309, 1955.

[6] K. J. Arrow, *Social Choice and Individual Values*, Wiley, 1951.