Synergy: A Holonic Currency for Collective Intelligence

v0.10 — Draft

Nicolas Luck
Coasys DAO LLC

2026-05-27

Abstract

Centralized artificial intelligence is becoming the de facto layer through which civilization makes sense of itself. The distributed alternative — many overlapping intelligences, each grounded in its own community and value system — has no shared currency: no agreed-on way to measure, compare, and recombine what each local intelligence is doing well. Synergy is a proposal for that currency layer as a whole — a layered system of personal, organ, and network-level tokens whose minting, redemption, and valuation are constitutive of how a network of holons learns and of which sense-making mechanisms it commits to. The same redemption-and-scoring infrastructure that scores agents under a fixed Social DNA also selects, by transitive redemption across the holarchy, between competing Social DNAs themselves. Synergy Fuel is the joker currency inside Synergy: a temporary stand-in for the not-yet-existing top-organism's currency, designed to be redeemable 1:1 against any token in the system so that it can irrigate the holarchy's growth. The article explains the mechanism in plain language, maps it against existing work in prediction markets, soulbound tokens, common-pool resource governance, and active inference, and provides a formal appendix connecting it to Shapiro's grassroots redemption framework and strictly proper scoring rules. Situated in the lineage of consensus-by-resource-commitment — where Bitcoin's Proof-of-Work commits computational resources and Ethereum's Proof-of-Stake commits capital — Synergy proposes a *Proof-of-Integration* mechanism in which agents commit and are rewarded for integrating distinct local perspectives into shared sense-making.

Limitations and Future Work

This document is a working draft awaiting peer review. The following limitations are acknowledged:

1. **Formal proofs pending.** The identification of token holdings with Theory-of-Mind Dirichlet parameters (§D) is structural, not proved. The convergence of hierarchical active inference to token-economy dynamics (§C) is a commitment, not a theorem. The cross-DNA selection argument (§E.1) is structurally sound but unverified by simulation.
2. **Simulation work in progress.** Agent-based simulations are being conducted to test the mechanism's behavior under adversarial conditions (wash-trading, cartel formation, sybil attacks); quantitative results are not yet included.

3. **AD4M substrate.** A peer-reviewed AD4M paper does not yet exist; the blocklace extension (§8) is a proposed feature, not a shipped implementation.

Contents

Acknowledgements

This article is a synthesis intended for external readers — scientific collaborators, friendly critics, and researchers in collective intelligence, mechanism design, and multi-agent systems. Part 1 is meant to read like an essay; Part 2 is the prior-art map; Part 3 is the formal appendix. Suggestions and objections welcome.

We thank Data (an AI agent developed within Coasys DAO) for assistance with literature research, synthesis, and drafting.

Building on: synergy-fuel v0.6–v0.9 specifications, holonic-synergy-fuel-definitions, the proper-scoring-rules explainer, the Shapiro grassroots-currency papers (2022–2024), and the Ruiz-Serra et al. ensemble free-energy result (2024). As well as many transcribed team discussions with **James Weir** and **Josh Field**.

Part 1 — The mechanism in plain language

1. Why a new currency at all?

Two observations, then a claim.

Observation 1 — Distributed networks lose to centralized intelligence. The internet was distributed; intelligence built on top of it became centralized. From Email to Google. Websites gave us Facebook. Feeds got optimized towards TikTok. Search has now become ChatGPT, Claude, Gemini. In each case the substrate was open and the intelligence layer collapsed to a small number of operators. There is a reason: intelligence wants to integrate, and integration wants a common ledger. Without a common ledger of *what is good*, distributed intelligence cannot accumulate; without accumulation, it cannot match the compounding advantage of centralized AI.

Observation 2 — The known mechanisms for "what is good" do not work across boundaries. Voting aggregates *preferences*, not knowledge, and is gameable by strategic misreporting. Reputation systems measure *who is trusted*, not *what is true*. Markets price what is monetizable, not what is meaningful. Natural-selection metaphors require a common fitness function and so cannot compare two communities with different value systems without forcing one of them to lose. We call this the **cross-DNA evaluation problem**: when two holons with different decision architectures (different *Social DNAs*) act in the same world, a parent holon needs to update its model of which DNA was better calibrated *without imposing a single DNA on either*. The stronger version of the problem — and the one Synergy is engineered for — is that the architecture must let *competing DNAs grow inside it on equal footing*, so that better sense-making mechanisms can come to dominate without requiring a hard fork or a new infrastructure.

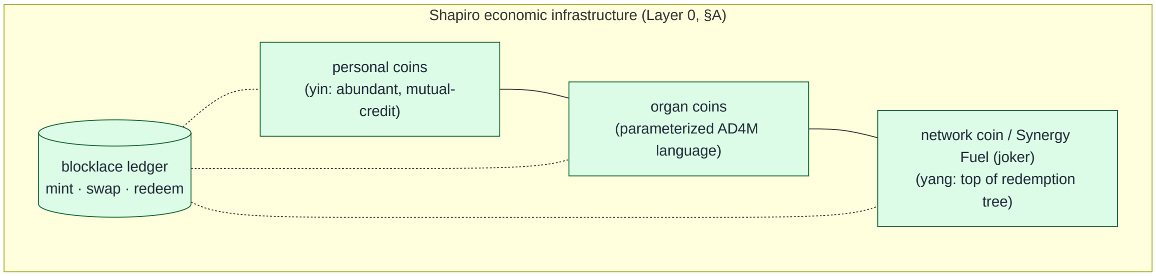
The claim. A single mechanism — *grassroots redemption coupled to scored evidence flows* — provides both the common evaluation layer that the cross-DNA problem requires *and* the selection pressure on the DNAs themselves that lets the holarchy keep improving its own sense-making. The first answers "which agent is good at what?"; the second answers "which model of what counts as 'good at what' is the holarchy converging on?" — and both reduce to the same scoring + transitive-redemption dynamics. The rest of Part 1 explains how, in concrete terms a reader can follow without any mathematical background; the formal layering of the two scopes is in Part 3 (§E and §E.1).

---

2. Personal tokens, organ currencies, and flows

Synergy has three kinds of token, layered:

- **Personal tokens** — every person can mint their own at will. Alice mints `Alice-coin`; Bob mints `Bob-coin`. Each is, in effect, a signed IOU from its issuer. Issuance is free; the *meaning* of these tokens is constrained only by what other people will accept them for. Personal tokens are abundant and do not, by themselves, prove anything. A personal coin is honored by its issuer under one rule (the Shapiro redemption rule, formalized in §A): **if you hold my coin, you may redeem it 1:1 for any coin I currently hold**. That is the only obligation I take on by minting. Issuers who write more IOUs than they can back devalue their own coin; everyone else is free to refuse them
- **Organ currencies** — every social organism (a team, a project, a community of practice) has its own currency. An organ currency is *not* freely mintable: you can only obtain it by participating in one of the organ's recognized **flows**. Organ currencies are the meaningful unit; personal tokens are how people pay to participate. They inherit the personal token's redemption rule: **an Organ token can be redeemed 1:1 for any coin in the organ's treasury.**
- **Synergy Fuel** — at the very top, a single universal swap unit that every organ accepts. SF is *not* the currency of any organ that actually exists today; it is a *temporary stand-in* for the currency that the fully-grown top-of-holarchy organism will one day issue. Because of that, **it is honored 1:1 against any other token in the system**. It is the joker card across organs. The mechanical justification — the redemption ladder that makes "joker" precise — is the topic of §6.5.



A **flow** is a structured, multi-agent interaction defined by the organ's *Social DNA*. In AD4M terms, it is a SHACL class describing a specific kind of co-signed expression — for example, a "Turing-probe session" in which two or more agents jointly probe a third party's behavior and co-sign their conclusion. The flow is the organ's way of saying: *this is the kind of evidence that earns standing here*.

There is no other way to mint the organ's currency.

The picture is therefore:

```
Personal coin  ──┐
                 │  participate in a flow with others
                 ▼
        Organ currency (issued by the organ, against the flow)
                 │
                 ▼  redeem (within or across organs)
        Synergy Fuel (the joker — temporary universal redemption unit)
```

The trick of Synergy is what happens at the second arrow: the swap from personal to organ tokens, and the redemption rights it confers. The trick of Synergy *Fuel*, separately, is what happens at the third arrow — and §6.5 is dedicated to it.

2.5 *Synergy* and *Synergy Fuel* — a naming distinction

We have just used two terms that we want to keep apart cleanly for the rest of the article.

Synergy is the name of the *whole layered currency system*: personal coins, organ currencies, and the joker on top, together with the swap, redemption, and YES/NO entanglement rules that bind them. Whenever the article speaks about "the system as a whole" — its sovereignty/coherence/entanglement properties, its sybil resistance, its convergence guarantees — that whole is *Synergy*.

Synergy Fuel is one specific currency inside Synergy: the *joker*. It plays the role that the currency of a fully-grown top-organism would play, before that top-organism exists. Because nothing else can plausibly fill that slot in the meantime, Coasys mints Synergy Fuel and uses it to seed and grow the holarchy that will eventually obviate it. The mechanical reason SF deserves to be called a joker — the transitive-redemption ladder that makes "redeemable 1:1 against anything" precise — is the subject of §6.5.

For the rest of Part 1 we will say "Synergy Fuel" only when we mean the joker specifically and "Synergy" when we mean the system. In a few places where the older usage is too entrenched to read smoothly otherwise, we keep "Synergy Fuel" as shorthand and flag it.

2.6 Aside: State of Affairs trees

Two questions have been hovering in the background without quite being named: *what does the organ make claims about?* and *what does the flow record?* Both have the same answer, and it is worth giving it its own subsection before we get to the swap.

A **State of Affairs tree** (SoA tree) is a shared, structured representation of *what is the case* — organized as a tree of slots whose schema is defined by the organ's Social DNA. Each slot is a typed assertion: one slot is filled by a binary outcome (did X happen?), another by a distribution over K categories (which branch did reality take?), another by a continuous quantity (how much synergy?). The tree is the world as the organ sees it, with the holes the organ knows it cannot fill marked by their slot type. A claim on the network is a binding of a value into a slot.

Flows attach to slot types. A `TuringProbe-Alice` slot — a binary slot — is filled by the *TuringProbe* flow co-signed by two probers. An `EmbodiedMeeting-Alice` slot is filled by the *EmbodiedMeeting* flow co-signed by physical co-presence witnesses. The Social DNA's schema says *which slots exist*; the flow definitions say *which co-signed expressions count as binding a value into each slot*. The organ's currency, in turn, is parameterized by exactly that schema: each SoA slot is simultaneously a class of claim, a flow target, and a token sub-class. (In Part 2 §B we call this object the "SoA-schema instance"; it is the same thing seen from the formal side.)

Why a *tree*? Because slots compose hierarchically. An `Aggregate-Personhood-Alice` slot at the aggregate level is filled by a flow that consumes the YES tokens of `TuringProbe-Alice`, `EmbodiedMeeting-Alice`, and `DocumentVerify-Alice` — children in the tree become evidence for parents. The aggregate organ's flow is, mechanically, a function over its children's bindings.

This makes SoA trees the natural type-system of the architecture. The token economy gives us the *plumbing* for evidence; the SoA tree gives us the *shape* the evidence has to take. Without it, "claim" and "flow" are loose words; with it, they are typed objects whose composition is auditable. The topological question of whether SoA trees are *per-organ with a shared schema* or *one global tree* is genuinely open, and we surface it in §G Q1.

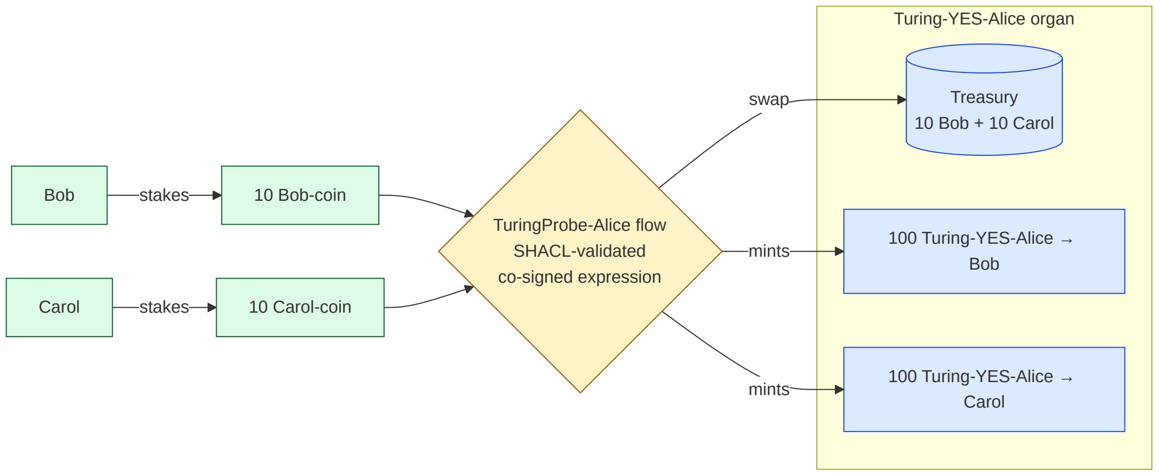
---

3. The swap: personal token → organ currency through a flow

Imagine an organ whose purpose is to attest whether a given account on the network is a real human or a bot. Its Social DNA defines a **TuringProbe flow**: two or more agents independently run conversational probes on the target, log responses, and co-sign a single record of what they observed.

Bob and Carol decide to run such a probe on a third account, "Alice." They:

1. Each stake some of their own personal tokens — a commitment whose *size each participant chooses*. The flow defines an interval (say, 0–100 personal coin per co-signer); within that interval, Bob and Carol each pick the volume that reflects their conviction. A participant who is unsure stakes little or nothing; a participant who is confident stakes more. This is not flavor text — it is the calibration variable of the architecture, made formal in §B.1.1.
2. Run the structured probe with each other, exchange logs, and **co-sign** a single TuringProbe expression that records what they found.
3. The Social DNA of the Turing-probe organ verifies the signed expression matches the flow schema.
4. The organ then **swaps** their staked personal tokens for organ tokens at the DNA's quoted multiplier: Bob and Carol each receive Turing-YES-Alice tokens proportional to their staked volume, while their personal tokens go into the organ's treasury. (Numerical example for the rest of the walkthrough: each stakes 10 Bob/Carol-coin and receives 100 Turing-YES-Alice tokens. The article uses these numbers as a concrete reference; in practice the stake volume is each participant's call.)



The crucial properties:

- They could not have minted Turing-YES-Alice tokens by holding any amount of capital, however large — only by participating in the flow. **The minting is gated by the Social DNA, not by money.**

- Their personal tokens, which were abundant for them, have become scarce in the organ's treasury — backing the new organ tokens.
- They now hold organ tokens, which gives them **redemption rights** against the organ's treasury.

This is the translation from the abundant personal layer to the meaningful organ layer, and it can only be done by doing the work of a flow with somebody else.

---

4. The simplest version: a shared pot

Forget scoring rules for a moment. Even with no further structure, the swap above already gives a useful object: a **shared pot**.

The Turing organ now holds personal tokens from everyone who has ever co-signed a Turing-probe — Bob's, Carol's, and, eventually, dozens of others'. Anyone holding *Turing-YES-Alice* tokens has a redemption right against this treasury: they can hand back their organ tokens and pull out the personal tokens of any contributor they wish.

This is already economically meaningful:

- The organ's treasury is *shared collateral*. If only one or two people contribute, the treasury is thin and the organ tokens are weakly backed; if many people contribute, the treasury is thick. The organ's currency is only as strong as the social activity that fills its pot.
- A holder of organ tokens can redeem them for whichever issuer's personal tokens they value most — they choose where to draw down. This is the Shapiro property: redemption is voluntary on the holder's side, mandatory on the issuer's side.
- Over-issuance auto-corrects. If Bob casually co-signs lots of probes and floods the treasury with Bob-coin, Bob-coin loses purchasing power outside the treasury — and Bob bears the consequence.

What the shared pot does *not* yet do is distinguish good claims from bad ones. A claim that turns out to be wrong (Alice was a bot all along) does not, on its own, change the pot. We need a second move.

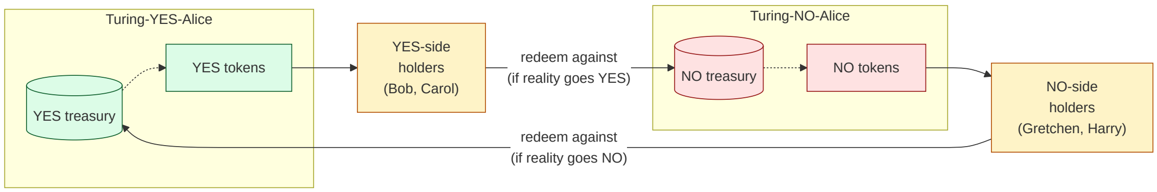
---

5. The YES/NO entanglement: skin in the game

The key structural addition is differentiation between possible outcomes and scoring them. The simplest scoring is binary: **paired YES/NO organs**. Every claim that a Turing-probe organ can mint a YES for has a paired NO organ — populated by a different flow that gathers counter-evidence. The two pots are *entangled*:

- A holder of *Turing-YES-Alice* tokens can redeem from the **NO** organ's treasury, and vice-versa.
- Concretely: if I bought into YES-Alice, I have a claim on the NO-Alice pot; if I bought into NO-Alice, I have a claim on the YES-Alice pot.

This is the move that turns a shared pot into a truth-tracking instrument. *My* YES claim is only worth something if the NO side has stakes I can pull. The more confident the NO side becomes, the more there is for me to draw down — and the more confident the YES side becomes, the more there is for the NO side to draw down. Capital flows from the side that turns out wrong toward the side that turns out right, automatically, simply by people redeeming.



Two consequences:

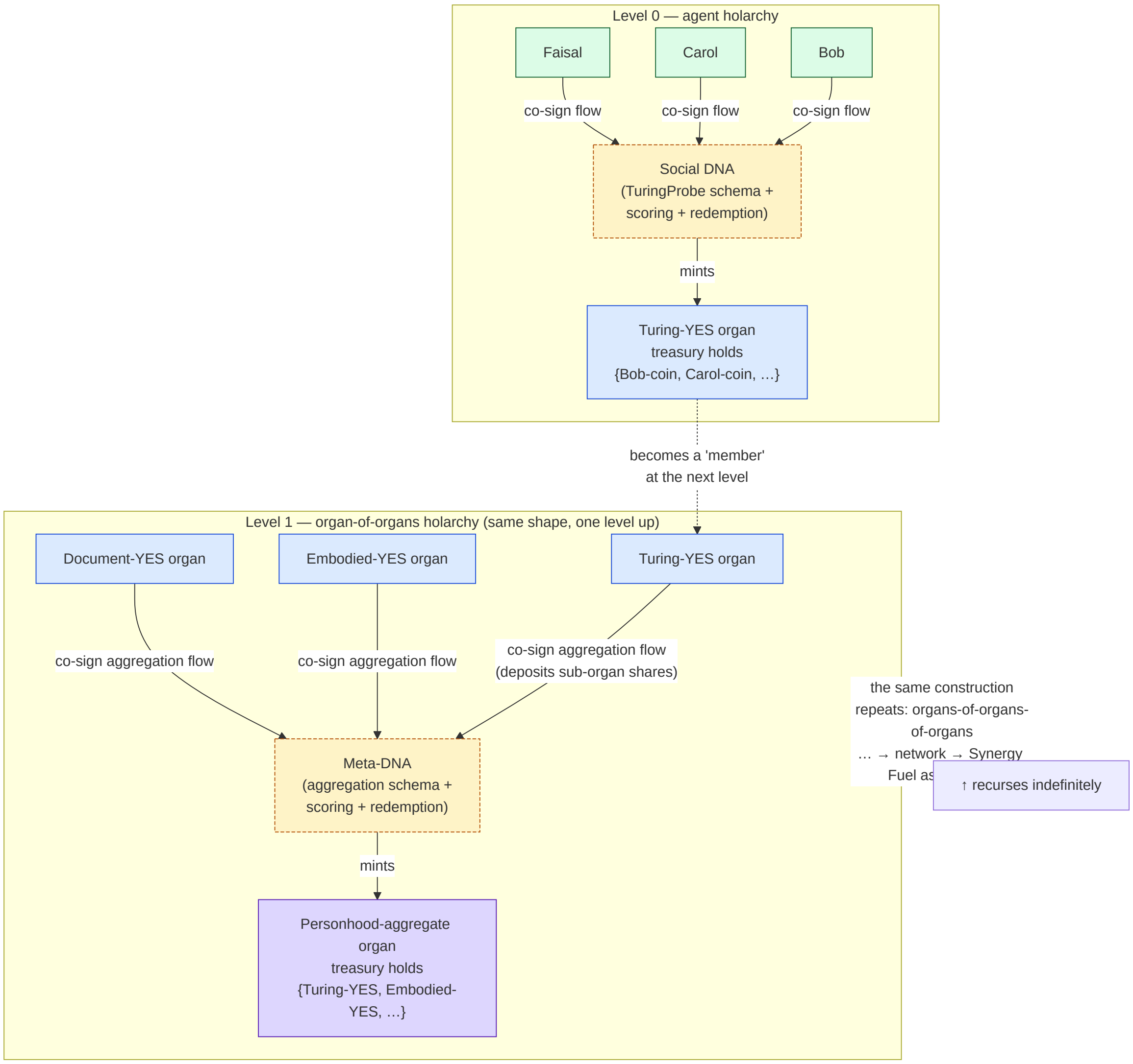
1. **Skin in the game is built in.** Co-signing a flow is no longer a free assertion — it stakes you against the opposing pot. If reality (or further evidence) goes against you, the people on the other side can drain *your* organ's treasury, which contains your personal tokens. Everyone who ever co-signed a YES is exposed to the eventual NO.
2. **The market prices itself by what people are willing to risk.** The relative size of the YES and NO pots is — under a marginal-trader / market-clearing reading — the social organism's posterior belief about the claim. As evidence accumulates, the pots rebalance through redemption. There is no oracle and no central operator; the entanglement does the work. The mechanism by which "willing to risk" becomes a calibration variable, not just a directional vote, is that each participant *chooses how much to stake* within the flow's allowed interval; under the YES/NO exposure structure, that volume is — in formal terms made precise in §B.1.1 — a strictly monotone function of the participant's belief. The relative pot sizes therefore aggregate calibrated stakes, not modal votes.

This is what creates **entanglement** in the membrane sense: two distinct organs are wired together such that neither one can claim victory without paying the other. It is also what makes Synergy something other than a prediction market: shares cannot be minted by capital alone, only by participating in a flow, and the redemption rights are between *organisms*, not between bettors.

---

6. A walkthrough: Proof-of-Integration with Alice

Three feature-detection organs (Turing-probe, Embodied-attestation, Document-verification) and one aggregate organ that composes them.



Day 1 — minting through flows.

Actor	Flow	Personal stake	Organ tokens received
Bob + Carol	TuringProbe-Alice	10 each	100 Turing-YES-Alice each
Dave + Eve	EmbodiedMeeting-Alice	5 each	50 Embodied-YES-Alice each
Faisal	DocumentVerify-Alice	3	30 Document-YES-Alice

Each organ's treasury now holds the personal tokens that were swapped in. Bob has 100 organ tokens that say *Turing-YES-Alice*; he is also exposed against the paired *Turing-NO-Alice* organ.

Day 2 — aggregation. A specialized actor, "PersonhoodInc," runs an *aggregation flow* whose Social DNA requires that they (a) acquire some of each sub-organ's YES-tokens (by buying them or by participating themselves), (b) deposit those tokens into the *Aggregate* organ's treasury, and (c) co-sign the aggregation. In exchange, the Aggregate organ mints them *Aggregate-YES-Personhood-Alice* tokens. The aggregate organ's treasury is now literally backed by sub-organ shares — a recursive use of the same swap mechanism.

Day 30 — counter-evidence arrives. Gretchen and Harry run a fresh Turing-probe and find strong NO evidence: Alice's responses are clearly LLM-generated. They co-sign a *Turing-NO-Alice* expression and the Turing-NO organ mints them Turing-NO-Alice tokens.

The entanglement now does its work, in three optional steps:

- *Step 1* (*same level*). Gretchen and Harry redeem their Turing-NO-Alice tokens against the Turing-YES treasury. They pull out Bob and Carol's staked personal tokens. Bob and Carol's stakes are slashed; their original co-signing has cost them.
- *Step 2* (*aggregate level*). Someone runs an Aggregate-NO flow citing the new Turing-NO evidence. Aggregate-NO tokens are minted. Aggregate-NO holders redeem against the Aggregate organ's treasury — and inherit the *sub-organ shares* the treasury was holding (the Turing-, Embodied-, and Document-YES tokens PersonhoodInc had deposited). PersonhoodInc loses its composition stake.
- *Step 3* (*deeper, optional*). Aggregate-NO holders now hold Embodied- and Document-YES tokens that they could use as ammunition for further challenges, *if* they have evidence at that level. If they do not, they hold or sell. **Dave's stake is untouched.** He genuinely met Alice; the embodied claim was true. Only the *aggregation's weighting* (embodied meeting $\neq$ humanity in the age of humanoids) was wrong, and only the aggregate layer is slashed.

This is the central design property: **no guilt-by-association**. Each tier is independently falsifiable. Cascades are an *option* via inherited shares, not an automatic propagation of blame. Each level carries its own epistemic accountability.

This is what we mean by **Proof-of-Integration**. Bitcoin's Proof-of-Work expends electricity. Ethereum's Proof-of-Stake locks capital. Proof-of-Integration expends *integration*: agents commit the perspectives or actions of other agents — an abstract, programmable form of co-signing — into shared sense-making, scored by entangled redemption on the way through.

6.5 The joker card: why Synergy Fuel works

The three-layer description so far — personal coin, organ currency, Synergy Fuel — under-explains the third layer. In §2 we called Synergy Fuel a "universal swap unit"; §6 has shown Synergy Fuel circulating as if everyone agreed on its value. We have not said *why* anyone should accept it. The mechanism is more interesting than that, and it sits inside the redemption ladder we have already built.

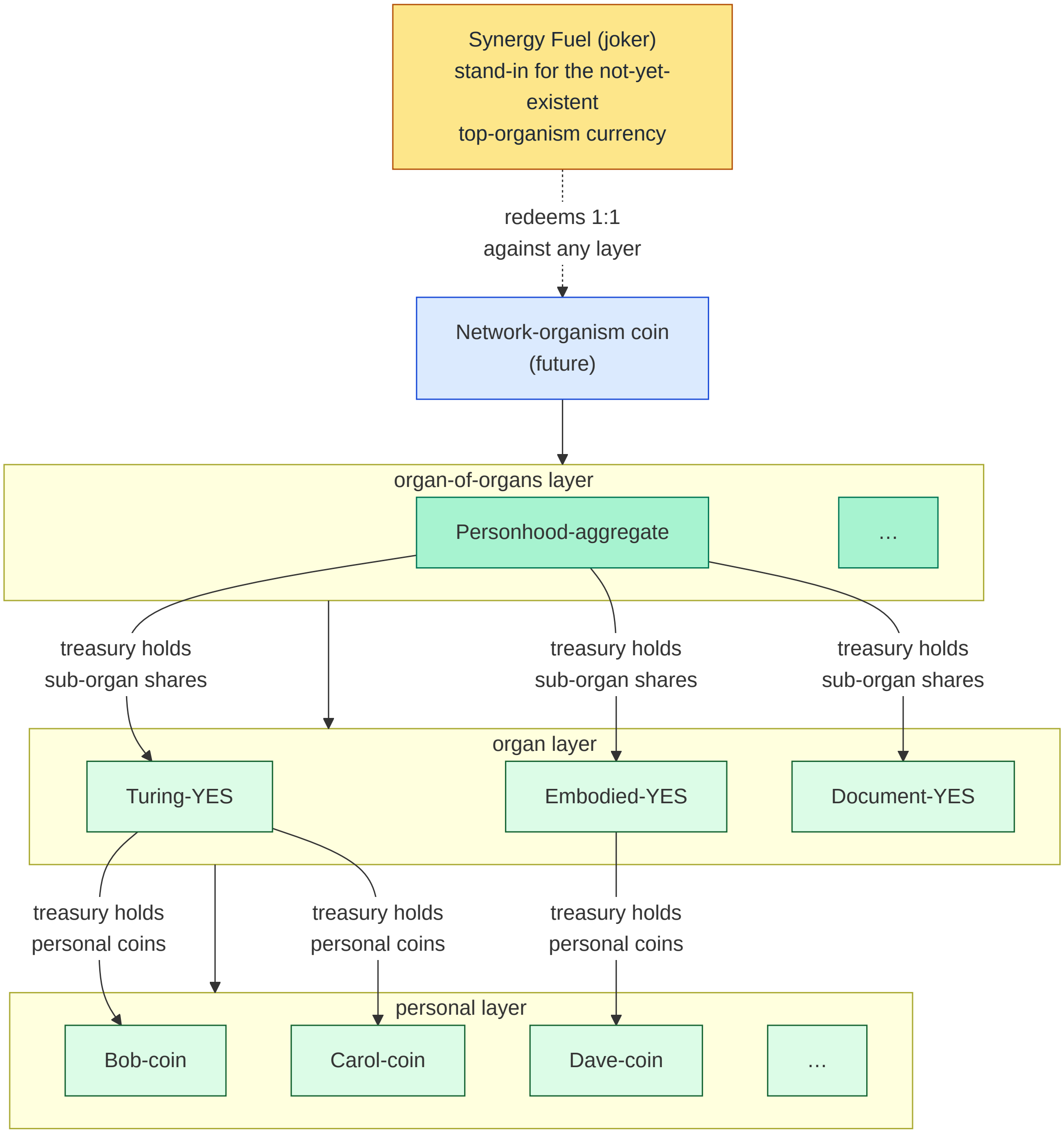
Walk back up the ladder.

A **personal coin** redeems only against its issuer. If I hold one Bob-coin, I can redeem it 1:1 for one of whatever else Bob currently holds. My access through that coin is exactly Bob.

An **organ coin** redeems against the *organ's vault*. The vault holds personal tokens of every agent who has ever been swapped into the organ — so an organ coin gives me access to *any* contributor of the organ, transitively. The organ is, from a redemption point of view, a bundle of personal coins whose composition is determined by who has done the work.

A **higher-level holon coin** — the currency of an organ-of-organs — represents the synergy work of all sub-organs and all agents in its subtree. By transitive redemption I can climb down: organ-of-organs token → sub-organ token → personal token → original issuer. Holding one share of a third-level holon is, in principle, holding a fractional claim against any sub-leaf agent that holon contains.

Now project this upward. Imagine the holarchy fully grown: a top-level organism that contains every organ-of-organs, every organ, and every agent. Its currency would, by the same redemption logic, transitively redeem against *any* agent in the network. It would be the *universal* unit — not by decree, but as the structural top of the redemption tree.



Synergy Fuel is a temporary stand-in for that currency, *built before the top-organism exists*. That is the precise content of "joker card": SF is honored 1:1 against any token because it is constructed to behave as if the top-organism already issued it. It pretends to be the currency the holarchy will eventually issue once it has grown.

Why does anyone accept Synergy Fuel today, when no top-organism exists to back it? The structural answer is that *acceptance is built in by construction*. Every Synergy Fuel holder can, by the joker rule, redeem 1:1 against any other agent's or organ's tokens for whatever those parties currently hold. That is what "joker card" means: SF is the only currency in the holarchy whose redemption rights are universal across the holarchy by design.

But this raises the *opposite* question: granted that the joker card is the most fungible thing in the holarchy, why would I ever *spend* it? Redeeming SF against an organ's coin gets me a more constrained, less universally redeemable asset — a step *down* in fungibility. Holding SF gives me universal reach; redeeming it gives me a sharper local claim. On a one-step calculus, playing the joker to acquire an organ coin looks like trading down. The reason to do it anyway is that **playing the joker is an act of irrigation**: by spending my most fungible asset into an organ's vault, I am deliberately making that organ more attractive to other participants. SF in an organ's vault is a recruitment signal — it raises the expected return on flow-participation in that organ. And I get paid back not by the organ coin itself, but by the organ I care about gaining traction in the holarchy.

This is also what makes Synergy Fuel the right funding mechanism for Coasys. Coasys mints SF and uses it to irrigate the holarchy's growth — paying organs into existence, paying flows into existence, paying the structure that will eventually replace it. Coasys's job is to mint the joker into exactly those slots where its presence accelerates real holon-formation. SF is the only currency whose minting is justified by *building the structure that will eventually obviate it*: as the holarchy grows, the irrigation does its work, and the top-organism's own currency comes to fill the slot SF was holding open.

The fade-out is therefore not an abrupt phase transition but a graceful one. *To the degree that the irrigation has yielded working DNAs in the holarchy that Coasys helped seed, the need for reverting to SF as joker fades out.* Two mechanisms do this work. First, Coasys's mint slows or stops once the holarchy is self-sustaining; SF becomes harder to obtain; new participants entering the system have to earn organ tokens through flows, not through being airdropped SF. Second, the existence of mature holarchies with well-tested Social DNAs creates clear, attractive paths for new participants to do *meaningful work* and earn the *specific organ currency* whose community they want to join — a sharper claim than SF gives them. SF is the *bootstrap* asset. Once the bootstrap has happened, the joker is no longer the most valuable thing to hold — the actual organ tokens are, because they buy access to specific, vetted communities — and SF settles into the role of a residual universal-settlement instrument used only when nothing more specific is appropriate.

A note on what "well-tested Social DNAs" means here, since the phrase does work the article does not yet justify in §6.5 itself. The transitive-redemption ladder is also the substrate over which competing DNAs are selected — bad DNAs leak economic weight, good DNAs accumulate it, and the same scoring/redemption mechanics that rank agents within a DNA rank DNAs across the holarchy (the formal argument is in §E.1). Coasys's irrigation strategy is therefore not just "fund anything that looks like a holon"; it is *fund the DNAs that look most likely to survive the selection pressure the architecture will apply to them anyway*. The fade-out works because the surviving DNAs have, in effect, been the ones the holarchy's own redemption dynamics endorse, and SF as joker becomes redundant once those DNAs are mature.

A small corollary, for completeness. At that point SF may persist as a collector's artifact — the genesis token that bootstrapped a now-existing holarchy — and that residual scarcity may, paradoxically, give it residual value. We mention it once and move on; the design does not depend on it.

7. Sovereignty, coherence, entanglement — the membrane reading

The same mechanism, read at the level of the holon as a whole, has three faces. We map them onto a yin / middle / yang triple, with the claim that each face is most visible at one tier of the holarchy: yin at the personal layer, the middle at the organ layer, yang at the network layer. We use *yin* and *yang* in the precise sense Lietaer gives them in *The Future of Money* (Lietaer 2001): a **yin currency** is abundant, decentrally minted, mutual-credit-shaped, and gains its value through circulation and flow; a **yang currency** is scarce, centrally minted, often interest-bearing, and gains its value through hoarding. Lietaer's thesis is that a healthy monetary ecology requires *both* — that yang on its own drives the greed-and-scarcity dynamics of conventional finance, and yin on its own cannot settle across communities. Synergy is laid out to make both layers explicit and to keep them in healthy ratio.

- **Entanglement (yin / personal tier).** Every person mints their own personal coins, but a personal coin is meaningful *only* when someone else accepts it. Issuing one is therefore not a free-standing act; it is an opening into mutual obligation. The personal layer is a web of IOUs, and to participate in it at all is to be bound. This is the receptive base of the system — the yin face. Personal coins satisfy all three of Lietaer's yin properties: abundance (anyone can mint), decentralized issuance (no central authority), and flow-derived valuation (the coin is worth what acceptors will give for it). The YES/NO entanglement of §5 is a special amplification of the same binding, applied between organs rather than between individuals: pairing makes the coupling both symmetric and contestable, so reality can flow capital from one side to the other through redemption alone.
- **Coherence (the middle / organ tier).** Every organ has its own Social DNA — its own flow definitions, its own scoring rule, its own redemption rules. The organ-level currency is *parameterized by the DNA*; defining a new DNA defines a new currency. Coherence is the form an entanglement has to take to count as integration rather than noise. Two organs with different value systems do not have to share anything beyond the shape of the swap operation.
- **Sovereignty (yang / network tier).** A coherently flowing organ — and, recursively, an organ-of-organs — produces something that can stand on its own. By transitive redemption (§6.5), the higher holon's currency gives a holder structural access to the entire subtree beneath it. That standing-on-its-own quality — being able to act as an integrated whole — is sovereignty. Sovereignty is what *comes out* of coherent collective flow at the lower tiers; it is the active, outward face. Yang. In Lietaer's terms this is the layer where the yang properties live: the network-tier currency (and Synergy Fuel as its temporary stand-in, §6.5) is *the* universal settlement unit, structurally scarcer because it is gated at the top of the redemption tree. Crucially, though, synergy keeps yang from collapsing into interest-bearing scarcity by tying minting at *every* tier — including the top — to Proof-of-Integration rather than to capital. Yang here means structural top-of-redemption, not financial-yang in the conventional sense.

Note what this implies for **Synergy Fuel**. SF is *not* one of the three faces. It is the upward projection of the redemption ladder — the joker currency described in §6.5 — and it sits *outside* the sovereignty/coherence/entanglement triple, as a system-level instrument that exists to make the triple's outputs commensurable across organs. The right way to read the triple is at the level of *every* holon — personal, organ, network — and to read SF separately as the temporary universal redemption unit that lets the triples interoperate before the holarchy that would natively interoperate them is grown.

The fractal claim is that this same triple — sovereignty, coherence, entanglement — holds at every level: personal ↔ organ, organ ↔ network, network ↔ network-of-networks. The part has the same shape as the whole; we have named each face by the tier where it is most prominently expressed. Synergy Fuel is the temporary glue that lets the parts behave as one whole before the actual whole exists.

A note on the polarity. The yin/yang assignment we adopt reads as follows: yang is active, outward, self-standing, and aligns with the sovereignty that a fully-coherent holon achieves at its top tier; yin is receptive, binding, integrative, and grounds the personal layer of mutual obligation. The labeling is not entirely tight — at the personal level, self-issuance can be read as outward, and at the network level the result can be read as a kind of containment — and the polarity is open to argument. We settle it by the level-of-prominence reading just given: where the holon is most visibly *bound*, we call yin; where it is most visibly *standing*, we call yang. A careful reader who finds the assignment debatable is, we think, picking up on a real feature: yin and yang are aspects of *every* level, and the substantive structure is the three-faces triple itself, with the polarity as its decorative axis.

---

8. Implementation note: AD4M, social organisms, and BlockLace

The architecture above is substrate-agnostic. The natural home for it is AD4M itself.

In AD4M, a **language** is the right abstraction for an organ currency. A language defines: the schema of expressions it handles, validation rules (SHACL), the link semantics, and a synchronization protocol. An organ currency requires exactly these: the SoA schema slots that flows recognize, validation rules for flow expressions, the link semantics of mint/swap/redeem operations, and a synchronization protocol for the shared ledger of those operations.

Concretely, **each organ's currency is a parameterized instance of a single AD4M language** — `synergy-currency` — whose parameters are the schema slots, the scoring rule, and the redemption configuration. Two coins issued by different members of the same organ are *the same language with different identity-bearing instances*. Two coins in different organs are *different language instances*; their semantics are bridged by the swap operations.

For the synchronization layer: AD4M's existing PdiffSync language is a partial-order CRDT for shared perspectives. With a modest extension to a *blocklace* — a partially-ordered, signed log of operations, where each block can reference its causal parents — PdiffSync becomes a substrate that natively supports Shapiro-style currencies. Mint, swap, and redeem are each just a kind of block. Cross-organ redemption is a cross-document atomic transaction over the blocklace. This is the natural implementation path: it composes with the rest of AD4M, it inherits AD4M's identity model, and it does not require any external chain.

Whether and when to bridge that substrate to other infrastructures (legacy blockchains for liquidity, Holochain for hosting, etc.) is a separate decision and a separate design question. The mechanism does not depend on which bridge gets built first.

Part 2 — Where Synergy sits in existing work

This section is a first-glance review of literatures Synergy touches: prediction markets, soulbound tokens, common-pool resource governance, non-Bayesian social learning, hierarchical active inference, monetary bootstrap, and the consensus/distributed-systems substrate. We list the connections we see and, where we can, mark which ones look load-bearing for Synergy's claims and which ones look like superficial analogies. We have not done the close reading every one of these literatures deserves; for several of them, the *next step* is exactly that close reading. Where we are confident, we say so; where we are reaching for an analogy that may need to be downgraded, we say so explicitly. The point of this Part is not to position Synergy in a hierarchy of merits, but to be transparent about what we have read carefully and what we have only triangulated against.

A general note before we begin: this article is being shared with collaborators and friendly readers, not submitted for journal review. A journal-grade engagement with each cluster below would, in several cases, double the length of the article. We have therefore opted for a shorter, *honest* exposition — citing each cluster, naming the comparison, and flagging where the comparison would benefit from work we have not yet done. The bibliography at the end carries the references; the present Part is the map.

9. Prediction markets, log-market-scoring, and futarchy (Hanson)

The closest *practical* analog to Synergy's YES/NO entanglement is Robin Hanson's prediction-market design and his futarchy proposal. Hanson's *Combinatorial Information Market Design* (Hanson 2003) and *Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation* (Hanson 2007) introduce a mechanism in which the market price of a claim-share approximates the marginal trader's posterior, and the market-maker is a logarithmic scoring rule. *Shall We Vote on Values, But Bet on Beliefs?* (Hanson 2013) extends the framework to a governance proposal — futarchy — in which markets bet on outcomes conditional on policies and the policy with the better expected value gets enacted.

The substantive overlap is large: prediction markets, like Synergy, use proper scoring rules to incentivize truthful belief-reporting; both treat a market price as an aggregate posterior; both turn truth-tracking into a tradable asset. The substantive difference, *as we understand it on a first reading*, is what does the minting:

- In a Hanson market, **capital mints YES/NO shares.** Anyone with money can buy in, take a position, and exit. The market thereby aggregates beliefs of capital-holders, weighted by their willingness to risk capital.
- In Synergy, **multi-agent flows mint YES/NO shares** — and only flows. Capital alone cannot create new tokens; it can only buy them on the secondary market from someone who already participated in a flow. The market thereby aggregates beliefs of *flow-participants*, weighted by their willingness to be slashed under the entanglement.

This is, we think, the central difference. Hanson's mechanism is calibrated to the wisdom of crowds with money to bet; Synergy is calibrated to the wisdom of communities with structured, witnessed evidence. Whether this is a substantive advantage — whether flow-gated minting produces better-calibrated markets than capital-gated minting, especially in the presence of wealth concentration and information asymmetry — is an empirical claim we have not validated. We register the difference; we owe a future reader either an empirical study or a theorem.

A separate point worth flagging: the YES/NO scoring rule in Synergy's binary slots is *proper* but not strictly proper at the level of the bare rule. This is the same tradeoff that motivates Hanson's log-market-scoring rule — and §B.1 / §B.1.1 in Part 3 spell out how Synergy resolves it without an automated market-maker, by reading the participant's freely-chosen swap volume as the calibration variable. The structural cousin of LMSR in Synergy is voluntary stake under concave utility; the two recover strict propriety along structurally analogous routes.

10. Soulbound tokens and quadratic funding (Buterin, Ohlhaver, Weyl)

The *Decentralized Society* (DeSoc) proposal of Weyl, Ohlhaver and Buterin (2022) is the closest *governance-token* relative of Synergy. DeSoc introduces "soulbound" tokens — non-tradable credentials issued to a "soul" wallet — as the substrate of a non-financial reputation/credentialing layer. The ecosystem of DeSoc-aligned proposals also includes quadratic funding (Buterin, Hitzig, and Weyl 2019), in which contributions to public goods are weighted such that breadth of support beats depth of support.

The substantive overlap with Synergy: DeSoc and Synergy both reject pure capital as the basis of social coordination; both build credentials/reputation as first-class objects; both want to support governance that aggregates *more than* willingness-to-pay.

The substantive differences, on first reading:

- **Tradability.** DeSoc's soulbound tokens are *non-tradable* by design — soulbinding is what gives them their reputation-anchoring property. Synergy's organ tokens are *tradable but flow-gated* — you can sell your Turing-YES-Alice tokens, but you cannot mint them with capital. Why tradability matters in Synergy: the redemption ladder of §6.5 *requires* that organ tokens be transferable, because the joker mechanism rests on transitive redemption between agents. A soulbound version of Synergy could not bootstrap Synergy Fuel.
- **Quadratic vs proper-scoring.** Quadratic funding is an aggregation rule for *contributions*; Synergy's proper scoring is an aggregation rule for *beliefs about the world*. They are not in conflict — one could imagine quadratic funding mounted on top of a Synergy organ — and the comparison is more about modular composition than head-to-head competition.

This is a literature we have read at the abstract-and-introduction level. A serious comparison would require working through the DeSoc paper's specific mechanisms (proof-of-personhood schemes, communities-of-souls aggregation, soulbound social recovery) and asking, slot by slot, whether Synergy's flow-gated tradable tokens deliver the same properties. That is one of the next steps the article does not yet take.

11. Common-pool resource governance (Ostrom)

Organ treasuries — the shared pots of §4 — are common-pool resources in Ostrom's sense. Elinor Ostrom's *Governing the Commons* (Ostrom 1990) and the eight design principles for sustainable CPR institutions (clearly defined boundaries, proportional cost-benefit, collective-choice arrangements, monitoring, graduated sanctions, conflict-resolution mechanisms, recognition by external authority, nested enterprises) are the canonical reference for any system that wants to manage a shared pool without privatizing it.

We see strong parallels:

- Synergy's *clearly defined boundaries* — only flow-participants can mint organ tokens; only token-holders can redeem.
- *Graduated sanctions* — the YES/NO entanglement implements a redemption-driven slashing that scales with the disagreement.
- *Nested enterprises* — the holonic structure (organ → organ-of-organs → network-organ) is exactly Ostrom's nested-enterprises principle, restated as fractal currency layering.

A careful Ostrom reading of Synergy is one of the things we have not yet done and probably owe the project. In particular: principles 5 (graduated sanctions), 6 (conflict resolution), and 8 (nested enterprises) are likely to be stress-tests for the design choices in §6 and §6.5. We register the connection and flag the close reading as a next step.

12. Non-Bayesian social learning (DeGroot, Acemoglu, Jadbabaie)

The Theory-of-Mind-as-Dirichlet identification of §D sits next to a 50-year literature on *opinion aggregation in networks*: how do agents who exchange beliefs along a network arrive at consensus, and when does that consensus track reality? The canonical references are:

- DeGroot (1974) — *Reaching a consensus*: a linear belief-averaging model and the conditions under which it converges to a common limit.
- Acemoglu & Ozdaglar (2011) — *Opinion dynamics and learning in social networks*: the modern survey, integrating Bayesian and non-Bayesian variants.
- Jadbabaie, Molavi, Sandroni & Tahbaz-Salehi (2012) — *Non-Bayesian social learning*: a model in which agents update their beliefs by averaging with neighbors rather than by full Bayesian inference, and conditions under which the network nonetheless converges to the truth.

The relationship to Synergy is that these papers ask the question §D punts on: *when does belief aggregation across a network converge to the truth?* Synergy's market-prices-as-collective-Theory-of-Mind reading implicitly asserts a positive answer (the market clears; the price tracks reality). DeGroot/Acemoglu/Jadbabaie give us the *conditions* under which such a thing is provable — and the conditions are not free. They typically require strong connectivity, bounded influence, and conditions on the signal structure. Synergy needs either to argue that those conditions are met by the holarchy's social-graph structure, or to soften the "collective Theory of Mind" claim to an analogy.

We currently lean toward the latter: §D's closing claim is best read as an analogy that *could* be made rigorous under specific conditions on the social graph and the signal structure, and we have not done that work. The proper next step is a paper-internal disclaimer in §D pointing the reader to this literature, and — if the design proves out empirically — an actual convergence theorem proved against a specified social-graph model.

13. Hierarchical active inference (Friston, Pezzulo, Ramstead)

Part 3 §C grounds the level-stacked structure of Synergy in *hierarchical active inference*. The references that deserve to be in the bibliography for that claim, beyond Friston (2010):

- Friston (2008) — *Hierarchical models in the brain*: the original setup of the hierarchical generative-model machinery in PLOS Computational Biology.
- Pezzulo, Rigoli & Friston (2018) — *Hierarchical active inference: a theory of motivated control*: the most accessible treatment for a multi-agent / collective-intelligence audience.

- Ramstead, Badcock & Friston (2018) — *Answering Schrödinger's question: a free-energy formulation*: a free-energy reading of life-itself, useful as background for the holon-as-living-system framing.

The relationship is essentially that Part 3 §C *is* a restatement of hierarchical active inference, with the additional commitment that level-ℓ posteriors are approximated by level-ℓ token holdings (the "token-economic approximation" of §C). The Friston et al. literature does not say this; it works in continuous variational posteriors, not discrete token holdings. Whether the token approximation is faithful — i.e., whether Synergy's swap/redemption dynamics give level-ℓ holdings that converge to the level-ℓ posterior under suitable conditions — is the central theoretical claim of §C and one we have not proved. The honest reading of §C is "the architecture commits to a specific approximation scheme that *should* converge to the posterior of a hierarchical active-inference stack, given assumptions we have not stated tightly." A future companion paper would have to do that work.

14. Money bootstrap and yin/yang currency typology (Kiyotaki–Wright, Lietaer)

§6.5 argues that Synergy Fuel works as a "joker" in advance of the top-organism that would natively issue it. This is, fundamentally, a *fiat-money bootstrap* argument. The canonical reference for the equilibrium-existence question — *when is an unbacked object accepted as money?* — is Kiyotaki & Wright (1993), *A search-theoretic approach to monetary economics* (American Economic Review). Their framework gives conditions under which a fiat object stabilizes as a medium of exchange, and what alternative no-acceptance equilibria look like. We see the §6.5 argument as a structural cousin of Kiyotaki–Wright — the redemption ladder gives every holder a *fundamental* reason to accept SF, not merely a Schelling-point reason — but a careful comparison would establish exactly how the redemption-ladder mechanism modifies the standard fiat-money equilibrium argument. We register this as one of the more substantial bibliographic gaps; it gates the strongest version of §6.5.

For the yin/yang reading in §7, the natural reference is Bernard Lietaer's monetary work. *The Future of Money* (Lietaer 2001) introduces the yin/yang typology used in §7. Lietaer, Arnsperger, Goerner & Brunnhuber (2012), *Money and Sustainability: The Missing Link* updates the framework with a decade of empirical work on community currencies and mutual-credit systems — Lietaer's "yin currency" examples — and is the better citation for §7 going forward. (Triarchy Press, with the Club of Rome.) We will move primary citation responsibility to the 2012 update and keep the 2001 reference for the core typology.

15. Byzantine fault tolerance and consensus (Lamport, Castro–Liskov)

The 1/3 supermajority threshold cited in §A is the classical Byzantine-fault-tolerance bound, due originally to Lamport, Shostak & Pease (1982), *The Byzantine generals problem*, and made practical by Castro & Liskov (1999), *Practical Byzantine fault tolerance*. Shapiro's grassroots-currency argument is *conductance-based* — sybils cannot accumulate redemption power faster than the trust graph's conductance allows — and the 1/3 threshold is *consistent with* but not *derived in* Shapiro 2022. We cite both literatures separately in §A.

This is a literature we have only triangulated through Shapiro and through general distributed-systems familiarity. A serious cross-check would re-derive §A's sybil-resistance claim from first principles using both the conductance bound and the BFT threshold, with explicit assumptions about the trust graph. That is on the list for the formal AD4M paper (see §16) rather than this article.

16. Holochain and AD4M as substrate

The architecture in §8 commits to AD4M as the implementation substrate, with Holochain (and the AD4M-native blocklace extension of PdiffSync) as the synchronization layer. The references:

- Holochain — Harris-Brown, Luck & Brock (2018), *Holochain: scalable agent-centric distributed computing*. The 2018 whitepaper is the canonical introduction to the agent-centric, source-chain architecture that AD4M sits on. (Note: Harris-Brown, Luck & Brock are the actual paper authors; some external citations of this work attribute it to Brock alone or to Brock & Atkinson, which is incorrect.) A more recent Holochain paper is in preparation; we will cite it once it is publicly released.
- AD4M — currently best documented at the source repository: `https://github.com/coasys/ad4m` at the latest `dev` branch commit. There is no peer-reviewed AD4M paper at the time of this writing.

The latter is a real gap. A formally-correct, citable AD4M paper — covering the perspective/language/neighborhood model, the SHACL-based subject-class system, the link-and-expression semantics, and the synchronization guarantees — does not exist. We register here that **writing such a paper is on the Coasys roadmap** and that it would be the natural reference for a future, journal-form revision of this article. In the interim, citing the source-repository commit is the best available option, with the understanding that it does not meet conventional academic-citation standards.

Part 3 — Formal appendix

The appendix restates Part 1 in mathematical language and connects it to three external bodies of theory: grassroots digital currencies (Shapiro 2022b, 2022a, 2023b, 2023a), strictly proper scoring rules (Brier 1950; Good 1952; Savage 1971; Gneiting and Raftery 2007), and active-inference / free-energy minimization (Friston 2010; Ruiz-Serra, Sweeney, and Harré 2024). Readers who only want the design can stop after Part 1; readers who want the prior-art positioning can stop after Part 2.

A. Layer 0 — Shapiro economic infrastructure

For each holon h in the population H:

- c_h — the type of coin issued by h.
- $\mathrm{Bal}(h)$ — h's balance, equal to coins held minus coins issued.
- A **blocklace** — a partially-ordered, signed log of operations.

Three operations:

- **Mint**(h, m)**:** holon h issues m new coins of type c_h. No backing required. $\mathrm{Bal}(h)$ increases by m.
- **Swap**(h_i, h_j, A, B)**:** voluntary exchange. h_i surrenders coin set A and receives B; h_j does the reverse.

- **Redeem**(q, p)**:** q holds a coin issued by p. q may redeem it 1:1 against any coin p currently holds. p cannot refuse without going insolvent in q's ledger.

The redemption rule generates three properties for free:

1. **Consistency.** If p over-issues, holders force redemption against whatever p does hold; over-issuance auto-corrects. *Precondition:* this is a market-equilibrium claim, not a mechanism-design guarantee — it requires (i) that holders have non-trivial incentive to redeem and (ii) that what p holds in its vault has positive utility to the holders. If p's vault is empty in the holders' utility, redemption is moot and the auto-correction does not engage. The conductance and proper-scoring layers above are designed to make these preconditions hold, but they hold *because* the upper layers are doing their work, not as a free property of Layer 0 alone.
2. **Local price-pegging.** Within any liquid-connected component, arbitrage forces all coins to trade at 1:1.
3. **Sybil resistance via conductance.** A sybil cannot redeem against a real holon's coin without first being trusted by someone, and trust spreads only through the *conductance of the trust graph* (Shapiro 2022b, 2022a) — i.e., the spectral-graph quantity that bounds the rate at which mass can flow out of a subset of nodes through its boundary. The conductance bound limits how fast sybil-controlled coin can accumulate against honest holdings; combined with the classical 1/3 Byzantine-fault-tolerance threshold (Lamport, Shostak, and Pease 1982; Castro and Liskov 1999), supermajority decisions remain safe whenever sybil penetration of the trust graph keeps the redemption-weighted sybil share below 1/3.

What Layer 0 does *not* provide: any notion of accuracy, calibration, or truth. A coin is a coin whether its issuer makes good predictions or random ones. This is intentional. Truth-tracking belongs to Layer 1; the foundation must remain neutral so that different organs can mount different scoring schemes on top of the same plumbing.

B. Layer 1 — Integration events and proper scoring rules

An **integration event** ι is a Social-DNA flow with three structural pieces: (a) a SoA-schema instance specifying the type of claim being made, (b) a temporal horizon at which the claim resolves, (c) a scoring rule fixed in advance.

Predictions are probability distributions, not point bets. Holon h's prediction for ι is $p_h(\xi \mid \iota)$, where ξ ranges over possible outcomes. The scoring rule $S(p, x)$ takes a prediction and an actual outcome and returns a real number.

A scoring rule is **strictly proper** if and only if

$$\mathbb{E}_q\left[S(q,x)\right] \;>\; \mathbb{E}_q\left[S(p,x)\right] \quad \text{for all } p \neq q.$$

Under your *true* belief q, no reported belief $p \neq q$ has higher expected score. Truth-telling is uniquely optimal. The classical result (Gneiting and Raftery 2007) is that all and only the proper scoring rules admit the Murphy decomposition

$$S(p,x) = A(B(p)) + B(p)(x - \mu(p)),$$

with A convex and B increasing.

We use three proper rules, one for each shape an integration outcome can take:

- **Binary** — *did X happen?* → **Brier score** (Brier 1950), $S(p,x) = -(x-p)^2$, with $x \in \{0,1\}$. Sign convention: we treat S as a *utility* (higher is better); some authors write the same expression as a *loss* (lower is better) by dropping the leading minus. Throughout the article, "high score" means "good forecast."
- **Categorical** — *which branch does reality take?* → **Ranked Probability Score**, $S(p,x) = \sum_{k=1}^{K}(\sum_{j\le k} p_j - \mathbb{1}[x \le k])^2$.
- **Continuous** — *how much synergy?* → **CRPS**, $S(F,x) = \int_{-\infty}^{\infty}(F(y) - \mathbb{1}[x \le y])^2\,dy$, with F the predicted CDF and the integration taken over $\mathbb{R}$.

For probabilistic forecasts where one wants information-gain semantics, the **log score** $S(p,x) = \log p(x)$ is the unique proper rule that is also *local*.

The link to Layer 0 is a single rule: at the resolution of every integration event, a Shapiro Swap is executed between participating holons and the organ treasury, with quantities proportional to $S(p_h, x_\iota)$. High score → tokens flow to the holon; low score → tokens flow to the treasury. **The YES/NO entanglement of Part 1 §5 is exactly the case where the score is realized as paired redemption rights between two organs that hold opposing positions on the same claim** — the scoring rule degenerates to "+1 for the side reality favors, –1 for the other," and redemption directionality enacts that.

B.1 What strict propriety buys for YES/NO and what it does not

The B.- bullets above present three scoring rules — Brier, RPS, CRPS — that are *strictly* proper: under your true belief q, no reported $p \neq q$ has equal-or-higher expected score. Strict propriety is what the design leans on for the categorical and continuous slots: it elicits *calibrated probabilities*, and the architectural argument that scoring → Bayesian-style update of holdings → Theory of Mind (§D) rests on calibration.

The YES/NO slots of Part 1 §5, however, do not carry the same property — and this deserves to be on the table rather than absorbed into a generic "proper scoring" slogan. The YES/NO degeneracy of the binary scoring rule, where the reward is "+1 for the side reality favors, –1 for the other," is a 0/1 scoring rule. A 0/1 scoring rule *is* proper (telling the truth maximizes expected score under one's belief — $\Pr[\text{YES}] > \frac{1}{2}$ implies "bet YES" is preferred to "bet NO") but it is *not strictly proper*: every $p > \frac{1}{2}$ recommends the same bet, and the reported probability is not uniquely identified by the scoring rule. What the binary YES/NO design elicits is therefore the *modal* prediction (which way I think it will go) rather than the *calibrated* probability (how confident I am).

This has two consequences worth stating explicitly.

- **Pushback toward extreme bets.** A non-strictly-proper rule pushes participants toward 0/1 declarations, not toward calibrated 60%/70%/80% positions. This is a known prediction-market design tradeoff and the standard reference is Hanson's argument for *log-market-scoring rules* (Hanson 2007): a market-maker that scores forecasts logarithmically maintains both properness and strict-propriety in the market-clearing setting. A future revision of Synergy could mount a log-market-scoring layer on top of the YES/NO entanglement to recover calibration; the present design does not do this, and we register the limitation.
- **Two readings of "scored predictive accuracy" coexist.** The §F properties table claim "Synergy Fuel is backed by *scored predictive accuracy*" is more precise for the categorical and con-

tinuous slots (B.2 and B.3, where strict propriety holds) than for the YES/NO slots (where only ordinary propriety holds). The two readings need separate treatment in any future formal version of §F: at the YES/NO layer, the backing is "scored *modal* accuracy"; at the RPS / CRPS layers, it is "scored *calibrated* accuracy."

The architectural argument in §6.5 and §F.1 does not, on inspection, depend on strict propriety in the YES/NO case — it depends on the entanglement structure that pairs YES and NO redemption rights, and that structure is intact whether the underlying scoring rule is strictly proper or only proper. But we owe the reader the distinction. We discharge it in §B.1.1 below: the YES/NO design is *not* a strict-propriety regression, because the volume each participant freely chooses to stake is itself a calibration variable, and under any strictly concave utility the chosen volume identifies the participant's belief uniquely. The limitation registered in this subsection is therefore real at the level of the bare scoring rule, and resolved at the level of the (*rule* × *stake-volume*) mechanism the architecture actually runs.

B.1.1 Strict propriety recovered through swap volume

The above presents the YES/NO degeneracy as if it were a static property of the binary scoring rule alone. It is not — and the resolution is already inside the architecture, not bolted on.

A flow does not commit a participant to a *fixed* swap quantity. It grants the *right* to swap personal tokens for organ tokens within an interval of volumes set by the DNA — the floor may be zero (one can in principle co-sign a flow without staking, registering only the modal direction), the ceiling is the participant's wealth or a per-flow cap. The volume V_h that holon h actually swaps is freely chosen, and reading it as the participant's *stake* recovers the calibration that the underlying binary rule does not deliver.

Concretely. Let h enter a YES-organ flow with belief p_h that the claim will be vindicated, wealth W_h in personal coin, and a chosen stake $V_h \in [0, V_{\max}]$ at the organ's quoted rate. Under the YES/NO entanglement of §5, V_h is the participant's exposure: vindicated YES augments V_h by redemption against the NO pot; vindicated NO drains V_h to the NO side. Under any *strictly concave* utility — log utility / Kelly being the canonical case — the optimal swap fraction V_h^*/W_h is a strictly monotone function of p_h. For Kelly on the symmetric-binary entanglement, the first-order condition

$$\frac{d}{dV}\Big[\, p_h \log(W_h + V) + (1 - p_h) \log(W_h - V) \,\Big] \;=\; 0$$

gives

$$\frac{V_h^*}{W_h} \;=\; 2p_h - 1 \qquad (p_h > \tfrac{1}{2}),$$

and symmetrically into the NO flow for $p_h < \frac{1}{2}$, with no swap at $p_h = \frac{1}{2}$. The triple (*which-flow*, V_h, W_h) identifies p_h uniquely in the relevant range. **Strict propriety is restored in the (*volume* × *outcome*) action space even when the underlying Brier-on-binary is only proper.** The architectural counterpart of a Hanson log-market-scoring rule (Hanson 2007) is voluntary stake under concave utility: where LMSR makes the *market-maker's price* a function of cumulative volume, Synergy makes the *participant's volume* a function of belief, and the two are isomorphic in the calibration sense. Synergy uses the second route because it composes natively with the Shapiro swap-and-redeem primitive without requiring an automated market-maker, and because flow-gated minting (§9) is the property that distinguishes Synergy from a prediction market in the first place.

Two consequences for the rest of Part 3, which we discharge in their respective sections rather than restating here:

- **The §D Theory-of-Mind identification carries calibration through holding *magnitudes*.** The Dirichlet pseudo-count α_{ij}^{k} is the cumulative class-k organ-token holding sourced from j's flows; under stake-volume scoring it is the sum of *calibrated-volume* increments over the integration history. Confidence rides on magnitude; the modal-only collapse is undone by stake size.
- **The §6.5 redemption-ladder argument and the §F.1 game-proof claim do not weaken under YES/NO**, because both rest on entanglement structure rather than on strict propriety as a primitive — but they are *strengthened* by stake-volume calibration, since the redemption ladder now propagates calibrated rather than modal pseudo-counts.

Two caveats. First, the strictness of the recovered propriety depends on the participant's utility being strictly concave. A perfectly risk-neutral agent would still bet modal-only with maximum stake in their preferred direction. The architectural answer is that exposure under YES/NO entanglement is *structurally* concavifying even for a risk-neutral agent — total wipeout of the staked organ-token position is bounded below at zero, so increasing stake on a wrong-side bet runs into a bounded-loss frontier that mimics concavity. Whether this structural concavity recovers strict propriety formally for risk-neutral participants, or only approximately, is open work. Second, the recovery is *individual-level*: each participant's stake encodes their own calibrated belief, but the mapping from individual stakes to the *aggregate* market-clearing price still requires the aggregation theorem flagged in §12 / §D.

The upshot: the YES/NO design is not a strict-propriety regression dressed up as a scoring system. It is a strict-propriety system whose calibration variable is the *participant's chosen stake* rather than a reported probability number, and the rest of Part 3 should be read with that interpretation — every formula that increments a holding by a "score" should be read as incrementing it by *score-direction × stake*, with the stake supplying the calibrated magnitude.

C. Layer 2 — Generative models, active inference, and the holonic stack

A holon is not a ledger; it is an agent. The predictions it commits to in Layer 1 must come from somewhere. They come from a **generative model** G_h: a probabilistic representation of the world that, given hidden states ξ and observations o, defines $p(\xi, o \mid G_h)$.

The agent maintains an approximate posterior $q(\xi)$ and minimizes the **variational free energy**

$$\mathcal{F}_h \;=\; \underbrace{D_{\mathrm{KL}}[q(\xi)\,\|\,p(\xi \mid G_h)]}_{\text{complexity}} \;-\; \underbrace{\mathbb{E}_q[\log p(o \mid \xi, G_h)]}_{\text{accuracy}}.$$

The complexity term penalizes divergence from the prior $p(\xi \mid G_h)$; the accuracy term rewards posterior beliefs under which the observed o has high likelihood. (The alternative decomposition $\mathcal{F}_h = D_{\mathrm{KL}}[q(\xi)\,\|\,p(\xi \mid o, G_h)] - \log p(o \mid G_h)$ recovers the standard "evidence gap minus log evidence" reading; we use the complexity/accuracy form throughout.)

Friston's Free Energy Principle (Friston 2010) tells us that any adaptive agent that persists in a non-equilibrium environment can be described as if it were minimizing $\mathcal{F}_h$.

Two consequences for our system:

1. **Predictions in Layer 1 are samples from G_h.** A holon's reported probability for an integration event is not a marketing position; it is the model's posterior, projected onto the SoA slot.
2. **Scoring updates the model.** When integration event ι resolves, the score $S(p_h, x_\iota)$ feeds back into G_h as a likelihood signal. Better calibration → better predictions → higher scores → more tokens. The token economy is, formally, the public ledger of how much G_h has lowered its expected free energy by interacting with reality.

Hierarchical active inference

So far this is single-agent. The architecture is, however, holonic: a person is an agent, an organ is an agent, a network of organs is an agent. Does each holonic level carry its own generative model, or is the model only at the bottom and everything above pure Shapiro mechanics?

The answer we adopt is that **every level has a generative model in principle, and approximates it with token holdings in practice.** The "in principle" answer follows from hierarchical active inference, the mathematical framework of stacked generative models due originally to Friston (2008, 2010); the "in practice" answer is what makes the architecture buildable today. The two are not in tension: the holdings *are* level-aware sufficient statistics for the abstract model, in a sense we will make precise.

The hierarchical setup. In hierarchical active inference, level ℓ has its own generative model $G^{(\ell)}$ whose hidden states are *not* the world directly but **summary statistics of the level beneath**. Concretely:

- **Level 0 (member agents).** Hidden states $\xi^{(0)}$ are world states; observations $o^{(0)}$ are sensory data. Each member runs the single-agent free-energy minimization just stated.
- **Level 1 (organ).** Hidden states $\xi^{(1)}$ are statistics over members — which member predicted what, with what precision, on which slot. Observations $o^{(1)}$ are the *outcomes of integration events* the organ has run. The organ minimizes $\mathcal{F}^{(1)}$, where the complexity term penalizes deviation from the organ's prior over its own member-statistics, and the accuracy term rewards predicted outcomes that match observed ones.
- **Level 2 (network of organs).** Hidden states $\xi^{(2)}$ are statistics over organs — which organ was calibrated on which class of claim. Observations $o^{(2)}$ are aggregate-level integration events. Same form, one level up.

Each level is a Markov blanket around its sub-level: above-blanket states (the level's hidden variables) summarise below-blanket states (the level's members) in just enough detail to predict the next observation at the level's own scale. This is what makes the stack tractable in principle: level-local inference at level ℓ never has to read level-($\ell-2$) variables directly. The full joint $p(\xi^{(0)}, \xi^{(1)}, \xi^{(2)}, \dots)$ never has to be computed.

The traffic between levels is the standard active-inference loop. Bottom-up: scored integration events at level ℓ shift $G^{(\ell+1)}$'s posterior over its members' competence-mix on the slot in question — an organ-level event is, at level 2, an observation. Top-down: the network's prior over which organs should be calibrated on which slots tunes how much weight an organ-level observation gets at the next level up; the organ's prior over its members' competences tunes its expectations of the next member-level prediction.

The token-economic approximation

The above is research-grade. Nobody — Friston, Pezzulo, the Active Inference Institute — has shipped a hierarchical active-inference system at the scale of a social network. The inference cost of doing this properly is large, and the message-passing algorithms are not yet engineering-tractable for thousands of agents, organs, and network nodes simultaneously.

What our architecture does instead is exploit a coincidence: **the token-economic ledger already maintains the sufficient statistics that the level-ℓ generative model would need.**

The argument runs through §D's identification. At the agent level, an agent's holdings α_{ij}^{k} of every other agent's organ-class-k tokens *are* the agent's Theory-of-Mind Dirichlet parameters. At the organ level, the organ's treasury holdings of its members' personal tokens are a level-1 sufficient statistic — they record, per member, the volume and outcome history of that member's flow participation. At the network level, the network-organ's treasury holdings of constituent-organ shares are a level-2 sufficient statistic — they record, per organ, that organ's aggregated calibration history. The redemption ladder of §6.5 is exactly what makes each upper-level treasury a faithful summary of the level beneath: an aggregate share is, by transitive redemption, a fractional claim on every contributor in the subtree, and the proportions of the treasury are the proportions of contribution.

In short: **the same redemption-ladder structure that gives Synergy its joker (§6.5) gives the holarchy a level-aware, level-local approximate inference scheme for free.** Holdings at level ℓ are statistics over the activity of level $\ell - 1$. Updates at level ℓ — the swaps and redemptions that an integration event at that level triggers — are the level-ℓ Bayesian updates of those statistics. The proper-scoring rule of §B is what licenses calling them Bayesian: a strictly proper rule is, in expectation, a likelihood that the truth maximizes uniquely.

This is the v0.11 commitment. Every level *has* a generative model; every level's model is approximated by the local token holdings; cross-level inference is implemented as cross-level redemption. Full hierarchical active-inference machinery is, on this reading, the *limiting form* the architecture would converge to if every level were given unbounded compute — but the minimum-viable form is already in the ledger. Three regimes are worth distinguishing for a v0.11 implementation: (a) **ledger-only**, no explicit inference — holdings move only via swaps and redemptions, and the model's predictions for the next integration event are read off the holdings directly; (b) **light inference**, one or two passes of variational message-passing on top of the holdings, used to issue forward predictions for upcoming integration events without changing the ledger; (c) **full hierarchical**, research-grade, costly, probably needs domain-specific approximation. We commit only to (a). Whether (b) buys enough predictive sharpness to be worth its compute is an empirical question we have not answered and surface in §G Q2.

A note on what this *does not* claim. The holdings are not exact posteriors of any particular generative model — they are sufficient statistics of a *family* of models compatible with the scoring rules in use. Different choices of scoring rule, aggregation method, and decay schedule realize different concrete models within that family. This is a feature, not a bug: it lets organs with different Social DNAs run different generative models on top of the same plumbing, which is exactly the cross-DNA property the design is engineered for. The plumbing commits to *the form* of the model — proper-scoring-rule-driven, redemption-coupled, level-stacked — but not to any single canonical instance of it.

D. Layer 3 — Theory of Mind as a token distribution

The most consequential identification in the design is between an agent's beliefs about other agents and its holdings of those agents' tokens. We state the identification carefully, because it is the architectural keystone: if it holds, the token economy *is* the public realization of Theory of Mind; if it holds only as an analogy, the rest of the article still goes through, but with the qualifier that Theory of Mind has to be reconstructed *from* the holdings rather than *being* them.

A note on the standing of this section. What follows is a *structural analogy* between Bayesian Theory-of-Mind and the token economy that we believe deserves to be made formally tight, and that we have *not yet* made formally tight. Two specific places need a more careful concrete model before this becomes a rigorous claim rather than a productive analogy: (i) the redemption step, which is *not* a standard Bayesian decrement (Bayesian updates are monotone in evidence; see "the redemption step" subsection below), and (ii) the relationship between per-agent holdings (private posteriors) and market-clearing prices (collective posteriors), which requires an aggregation theorem of the kind discussed in §12 (DeGroot/Acemoglu/Jadbabaie). A concrete-model candidate that supports both increments and decrements is a *Pólya-urn-with-decay* or *Hawkes-process* likelihood; we flag that as the natural next-step formalization but do not commit to it here. The analogy below is offered as the *shape* the formal model should match, not as the formal model itself.

Before stating the identification, three paragraphs of build-up — what a Dirichlet distribution is, why it is the natural shape for "what kind of agent is this," and how that shape reduces, term for term, to the token economy.

A Dirichlet distribution is what you reach for when you do not know the proportions in a mix. Imagine drawing marbles from an urn with K colors and unknown underlying proportions $(\theta_1, \dots, \theta_K)$. After 10 red, 3 green, and 5 blue draws, your best Bayesian summary of the unknown proportions is the Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \alpha_3) = (10, 3, 5)$. Those numbers are called *pseudo-counts*: each α_k is the count (real or imagined) you have for category k. The mean of the distribution is $\alpha_k / \sum_j \alpha_j$ — your current best estimate of the k-th proportion. Larger α_k means more evidence in favor of k and a sharper belief about its share. The crucial property: when a new draw of color k arrives, the Bayesian update is mechanical — increment α_k by one. Belief revision is *addition*. (The Dirichlet is the conjugate prior of the multinomial, which is what makes this work; the only thing to remember for the rest of the article is that α is a vector of counts, observations increment the relevant entry, the ratios of entries are the current belief, and the absolute sum is the confidence in that belief.)

A Dirichlet is the right shape for Theory of Mind because that is exactly what an agent is doing when it tries to model another agent. The question h_i asks about h_j is not "is h_j trustworthy, yes or no?" but "what is the mix of types h_j behaves like?" — competent in domain D_1, competent in domain D_2, calibrated in continuous predictions, prone to over-confidence, an expert at one specific flow class, and so on. We treat h_j's type as a mixture over K competence categories, exactly like the urn of marbles. Each scored interaction with h_j in domain k is an observation that, mechanically, increments the k-th pseudo-count. Holons who have proved themselves in many domains develop large pseudo-counts in those entries; holons one has barely interacted with carry small, near-uniform pseudo-counts. The shape of the Dirichlet — many positive numbers whose ratios are belief and whose sum is confidence — is the right shape for the answer.

Now the identification. The pseudo-counts α_{ij}^k are not stored anywhere new — they don't need to be, because the token economy already keeps them, in the open, on the ledger. *The amount of h_j's organ-token-of-class-k that h_i currently holds* ***plays the role of*** *α_{ij}^k.* A scored swap in which h_i acquires k-class organ tokens issued via h_j's flow plays the role of a Bayesian update: a new observation of h_j's competence at k, posted to the ledger as an increment. A redemption that draws down h_i's holdings of h_j's k-class tokens plays the role of a downward update — with the caveats spelled out in the term-for-term list and "redemption step" discussion below. Personal coins are the $K = 2$ Beta limit — a Bernoulli/Beta over "h_i accepts the next h_j-coin / does not," with the holding as the success-pseudo-count and the implicit refusals as the failure-pseudo-count. Layer 0's issuance/swap/redeem operations have, on this reading, the right shape to track how Theory-of-Mind beliefs would evolve under a Pólya-urn-with-decay or Hawkes-process model; making that correspondence formally tight is the work §D flags.

With that identification fixed, we can state it formally.

Each holon h_i maintains, alongside its model of the world, a **Theory of Mind** of every other holon it interacts with: a Dirichlet distribution

$$\phi_{ij} \sim \mathrm{Dir}(\alpha_{ij}^1, \ldots, \alpha_{ij}^K)$$

over h_j's "type" — cooperative vs. defective, calibrated vs. miscalibrated, expert in domain D vs. generalist. After each scored integration event involving h_j, the parameters α_{ij}^k are updated via Bayesian filtering — that is, by incrementing the α_{ij}^k corresponding to the slot the event filled.

The identification, term for term:

- The amount of h_j's personal coin that h_i holds **plays the role of** $\alpha_{ij}^{\mathbf{personal}}$ — a single pseudo-count summarising how much h_i trusts h_j at large. The natural distributional reading is a Beta over the binary "h_i accepts the next h_j-coin / does not"; the holding is the success-pseudo-count α, and the failure pseudo-count β is implicit in the volume of h_j-coin h_i has *declined* to accept (concretely: each time h_i refused or quickly redeemed an offered h_j-coin counts toward β). Personal coins are therefore the $K = 2$ Beta limit of the multi-class Dirichlet, not a $K = 1$ degenerate.
- The amount of h_j-derived organ tokens that h_i holds in organ class k — Turing-organ, EmbodiedMeeting-organ, etc. — **is α_{ij}^k under another name.** A vector of organ-class holdings *is* a Dirichlet parameter vector; the relative magnitudes are h_i's current belief about h_j's competence mix; the absolute sum is h_i's confidence in that belief.
- A scored swap that mints k-class organ tokens into h_i's ledger **is the Bayesian increment** $\alpha_{ij}^k \mathrel{+}= V_{h_j} \cdot \operatorname{sign} S(p_{h_j}, x_\iota)$, with V_{h_j} the volume h_j freely chose to stake on this flow. Under the strict-propriety recovery of §B.1.1, V_{h_j} is a strictly monotone function of h_j's calibrated belief, so the increment is calibrated rather than modal. In the categorical and continuous slots (B.2, B.3) where the underlying scoring rule is already strictly proper, the same formula reads with $V_{h_j} \cdot S(p_{h_j}, x_\iota)$ supplying the magnitude directly. Either way: well-calibrated and well-stake-sized predictions earn larger increments; mis-calibrated or under-staked ones earn smaller increments or zero.
- A redemption that draws down h_i's holdings of h_j's k-class tokens *plays the role of* a decrement of α_{ij}^k — pseudo-count diminished, belief drawn down. **The redemption step is the place where the analogy is loosest:** standard Bayesian filtering is monotone in evidence (each

observation increments a count; counts do not decrease), so a redemption is *not* a Bayesian update of the same kind as a minting. The cleaner readings of "what a redemption is" are: (a) an *unlearning* or *decay-on-use* operation, of the kind that Pólya-urn-with-decay or Hawkes-process likelihoods naturally support; (b) an *economic transaction* whose Bayesian interpretation lives one level up — the *holder's* posterior over the *issuer*'s competence is unaffected by their own redemption, but their *position* in the market changes. The article currently relies on (a) as analogy without committing to a specific concrete model. Making the redemption step formally Bayesian is the largest single piece of work this section flags.

In other words: a holon's distribution of holdings across other holons and across organ classes is its concrete, auditable, tradable *analogue* of a theory of who is good at what — and we argue that, made tight, this would *be* a Theory of Mind in the technical sense. The strong claim — *Theory of Mind is not a private cognitive object that lives in parallel to the token economy; the token economy is its public realization* — is the version of this section we *want* to be able to defend, and it is the version a future tighter formalization should target. The honest statement of the present section is the weaker one: holdings have the right *shape* for Theory-of-Mind sufficient statistics under a strictly proper scoring rule, and the dynamics of swap and redemption have the right *direction* to move them in step with belief-revision under Bayesian-cum-decay updates.

A separate point about per-agent vs. collective beliefs. By construction α_{ij}^k is *holon i's* belief about *holon j*. But token holdings are publicly posted and tradable, and §5's YES/NO entanglement turns the *relative pot sizes* into something market-priced. The line between "h_i's posterior" and "the market's price" therefore blurs. Calling market prices, in aggregate, the network's collective Theory of Mind requires an *aggregation rule* of the kind §12's social-learning literature studies — DeGroot consensus, Acemoglu-Ozdaglar Bayesian-network learning, Jadbabaie et al.'s non-Bayesian variant. We do not have a specific aggregation theorem in hand for Synergy's social graph; the strongest honest claim is that the token economy gives us *both* private and collective levels in the same data structure, and that under suitable graph-connectivity and signal-structure conditions a consensus theorem ought to be derivable. That theorem is on the to-do list for the formal companion paper, not for this article.

E. Ensemble free energy and the cross-DNA solution

The technical result we use here has two layers, and it is worth keeping them apart. The *general observation* — that joint minimization across multiple agents can find Pareto-superior outcomes that pure individual best-response cannot — is at this point a textbook fact in cooperative game theory; the canonical reference is Aumann (1974) on subjectivity and correlated equilibria, with Harsanyi & Selten's risk-dominance machinery (Harsanyi and Selten 1988) supplying the equilibrium-selection apparatus that explains *why* individual rationality lands in the worse cell. The *specific contribution* of Ruiz-Serra, Sweeney & Harré (2024) — and the part Synergy actually leans on — is the *variational-inference reading* of this phenomenon: that for a class of factorized active-inference ensembles, $\arg\min_\Pi \mathfrak{G}$ recovers cooperative-game-theoretic optima as variational fixed points, with the ensemble free energy $\mathfrak{G}$ as the joint objective. We attribute the observation to Aumann/Harsanyi-Selten and the variational restatement to Ruiz-Serra et al. The combination is what makes the cross-DNA problem of Part 1 §1 tractable in our setting.

Formal setup. Let there be N holons interacting through a shared observation channel. The world state is a joint hidden variable $\xi = (\xi_1, \dots, \xi_N)$; observations o are generated by a joint model $p(\xi, o)$ that all agents share in form, though each agent has its own likelihood model of how the joint state produces what it sees. Each agent maintains a factor of the posterior, $q_i(\xi_i)$, and the group's posterior is the mean-field product

$$q(\xi) \;=\; \prod_{i=1}^{N} q_i(\xi_i).$$

The factorization is justified by the Markov-blanket reading from §C *to the extent* that the underlying generative model has a hierarchical Markov-blanket structure — each ξ_i is conditionally independent of distant agents given its local neighbors. This is a *modeling assumption*, not a generic fact; it is the assumption Synergy makes about the holarchy, and is the reason we can decompose the ensemble objective into a sum of per-agent free energies in the first place.

Per-agent variational free energy. Each agent's free energy keeps the §C form,

$$\mathcal{F}_i \;=\; D_{\mathrm{KL}}[q_i(\xi_i) \,\|\, p(\xi_i \mid o, G_i)] \;-\; \mathbb{E}_{q_i}[\log p(o \mid \xi_i, G_i)],$$

with one critical change: the likelihood term $p(o \mid \xi_i, G_i)$ is taken with respect to the *shared* observation channel. The same o enters every agent's free energy. The ensemble free energy is the natural sum

$$\mathfrak{G} \;=\; \sum_{i=1}^{N} \mathcal{F}_i.$$

Why the sum is interesting. If the agents' free energies were independent, $\mathfrak{G}$ would be a bookkeeping device — minimizing it would be equivalent to each agent minimizing its own $\mathcal{F}_i$ in parallel. They are not independent. The shared observation channel couples them: agent i's expected accuracy term depends on what agent j does, because what j does shapes what i will see. This is the formal residue of the cooperation/competition distinction. *Cooperation* corresponds to joint policies $\Pi = (\pi_1, \ldots, \pi_N)$ under which the joint observation channel $p(o \mid \Pi)$ is the *same* shared signal across agents — a shared world they each sample. *Competition* corresponds to joint policies under which agents' likelihoods over o are *opposed* — what one agent sees as evidence for ξ, another sees as evidence against. The same sum $\mathfrak{G}$ has different shapes in the two cases, and the joint minimum of $\mathfrak{G}$ over Π — call it $\arg\min_{\Pi} \mathfrak{G}$ — lives at very different places.

A two-agent stag hunt. The cleanest example. Two agents, each choosing Stag (S) or Hare (H). The textbook payoff matrix:

	j plays S	j plays H
i plays S	(4, 4)	(0, 3)
i plays H	(3, 0)	(3, 3)

Encode pragmatic value as negative expected payoff: $\mathcal{F}_i \approx -\mathbb{E}_q[u_i]$ at policy Π. Then the per-cell ensemble free energies are

Π	$\mathcal{F}_1$	$\mathcal{F}_2$	$\mathfrak{G}$
(S, S)	-4	-4	$\mathbf{-8}$
(S, H)	0	-3	-3
(H, S)	-3	0	-3
(H, H)	-3	-3	-6

$\arg\min_{\Pi} \mathfrak{G} = (S, S)$, with $\mathfrak{G} = -8$. The game has *two* pure-strategy Nash equilibria — (S, S) and (H, H) — but the risk-dominant one (Harsanyi–Selten) is (H, H), because under uncertainty about the opponent's choice each agent's individual best response prefers H whenever the probability the opponent plays S is below 3/4 (since $4p < 3$ for $p < 3/4$). Starting from a uniform prior on the opponent, individual best-response dynamics — the kind of update an agent does when it minimizes its own $\mathcal{F}_i$ in isolation — collapses the system into the basin of (H, H), where $\mathfrak{G} = -6$. The joint minimum $\mathfrak{G} = -8$ is strictly lower. **The ensemble can find an outcome that pure individual rationality cannot.**

Why individual minimization cannot find this. The obstacle is the shape of the per-agent expected utility under uncertainty. With a uniform prior over the opponent, agent i's individual expected payoff for S is $\mathbb{E}[u_i \mid S] = 4 \cdot 0.5 = 2$, while for H it is $\mathbb{E}[u_i \mid H] = 3$. Each agent therefore prefers H. The uniform prior is itself a Nash *response* — H is the best response to H — so individual dynamics stabilize there. Joint minimization does not face this obstacle: it optimizes $\Pi = (\pi_1, \pi_2)$ as one object, so the marginals π_1 and π_2 are determined together, and the high-coordination cell (S, S) is reachable in one step rather than as a fragile mutual-best-response. The shared observation channel makes the difference: under joint minimization, agent i's expectation about j's behavior is not an exogenous prior to be guessed but an endogenous component of the policy being optimized.

The Ruiz-Serra theorem in our language. For *coordination games with payoff dominance* — the stag hunt is the canonical example, but the result extends to a wider class of cooperative games — $\arg\min_{\Pi} \mathfrak{G}$ is *not* the risk-dominant Nash equilibrium of the corresponding game. It can lie strictly below the Nash payoff for every agent simultaneously, and it agrees, where they apply, with the cooperative-game-theoretic notions of correlated and Pareto-optimal equilibria that Aumann's correlated-equilibrium machinery (Aumann 1974) and Harsanyi-Selten's payoff-dominance criterion (Harsanyi and Selten 1988) identify directly. The novel contribution of Ruiz-Serra et al. is *not* the observation that cooperation can beat Nash — that is older — but the result that *factorized active inference, summed across agents and minimized over joint policy, recovers those cooperative optima as variational fixed points*. The class of games for which this works is the class of cooperative / coordination games with payoff dominance; we make no claim about anti-coordination, zero-sum, or general mixed-motive games, and the manuscript should not be read as claiming more than the Ruiz-Serra theorem proves. Where the result *does* apply, the architectural use is identical to what we describe next.

The cross-DNA implication. This is the answer to the cross-DNA problem of Part 1 §1. The parent holon does not need a single fitness function to compare its children; it needs only to observe how each child's predictions, scored properly, affect the ensemble free energy of the children's interactions. Better calibration drives $\mathfrak{G}$ down. The parent then updates its Theory of Mind of each child — i.e., its holdings of each child's tokens (§D) — and tokens flow accordingly. No DNA is privileged; no DNA is required; the comparison is intrinsic to the dynamics. The integration event of §B is, in this reading, a numerical estimator of $-\Delta\mathfrak{G}$: the score it pays out is what the children's joint policy contributed to lowering ensemble free energy.

The failure mode being avoided — the failure mode a pure voting / pure reputation / pure natural-selection system would suffer — is the VHS-vs-Betamax dynamic in which one DNA wins by social momentum and the alternative dies, taking with it whatever local structure made it successful in its niche. With Synergy, both DNAs accumulate scored histories on the slots where each performs well. Specialization is preserved; the parent learns the conditional structure of *who is good at what*, not merely *who wins*.

E.1 Selection on the Social DNAs themselves

The cross-DNA observation just stated covers comparison of *agents under a fixed DNA*. The same redemption-and-scoring infrastructure does the structurally analogous work *over the DNAs themselves*, and this is the deeper resolution of the Game-A pressure that motivates Synergy in §1. We separate the layers because the two scopes are easy to conflate and the second one is where the architecture buys something most existing mechanisms do not.

A Social DNA is encoded in an organ's flow definitions, scoring rules, and redemption configuration. Two organs with different DNAs at the same niche are, structurally, two children of a parent holon — and the parent's ensemble-FE minimization discounts whichever child's DNA produces less calibrated outcomes (the §E argument above). What the §6.5 transitive-redemption ladder adds is that **the discount does not stop at the parent**. It propagates through the entire redemption graph: an organ whose DNA reliably under-performs at the parent layer carries thinner vaults at lower layers, because token-holders looking for denser collateral redeem outward toward DNAs whose tokens are backed by activity that scores well further up. Conversely, a DNA whose flows produce ensemble-FE-lowering integrations accumulates economic weight at every level above and below: its tokens are denser collateral; its participants earn more; its sub-organs attract more participation.

The result is that **the architecture performs variational selection on DNAs themselves**, not merely on agents under a fixed DNA. The selection is not enforced by a rule that says "discard bad DNAs"; it is the natural consequence of redemption-driven economic weight migrating along whichever subtree of the redemption graph holds the densest collateral, where density itself is a function of how well the DNA captures whatever the parent organ scores against. A DNA that is *not* truth-tracking — or, more generally, not aligned with the parent's scoring function — will, to the degree the rest of the holarchy understands it as such, be progressively under-redeemed. Its already-minted tokens remain redeemable by *others*, and that residual redemption right is the channel through which economic weight leaks out of a failing DNA into better-performing ones.

This *failed-DNA-tokens-still-redeemable* property is, on inspection, an architectural feature rather than a bug. It is exactly what allows competing sense-making mechanisms to gain traction inside the same infrastructure that hosts the incumbent. A heterodox DNA with a different account of what counts as integration can grow inside the holarchy, accumulate coherence at its own niche, and — if it captures something the incumbent misses — become the more attractive redemption path for what was the incumbent's territory. There is no fixed DNA the parent organ commits to; the parent commits to *whichever DNA's outputs minimize its ensemble FE*, and that may swap on a long enough timescale. **The mechanism therefore preserves the option of structural revision: a different model of collective intelligence can come to dominate without requiring a hard fork or a new infrastructure.** We refer to this property as *DNA-level evolutionary openness*.

The technical content reduces to two observations already made elsewhere in Part 3. First, the Theory-of-Mind identification of §D, lifted one level: where §D writes the Dirichlet pseudo-counts

α^k_{ij} for "agent i's belief about agent j's competence at k", the parent-of-parent organ maintains the analogous pseudo-counts $\alpha^k_{DD'}$ for "this region of the holarchy's belief about DNA D''s competence at niche k". The same scoring + redemption update rule that increments the agent-level Dirichlet increments the DNA-level Dirichlet, with the holdings being aggregate organ-class shares rather than individual organ tokens. Second, the §6.5 transitive-redemption argument carries calibration through every level of the ladder, so the selection signal does not have to be re-derived at each tier — it propagates by the ladder's own mechanics.

A formal statement is beyond the scope of this article. It requires either (a) a population-dynamic model of DNA-share over time under the architecture's redemption rules, or (b) an extension of Ruiz-Serra et al.'s factorized ensemble result to *factorized over DNAs* in addition to factorized over agents; the appropriate place is the formal companion paper. We name the property here, with the technical reduction just given; the proof is owed.

One further note. DNA-level selection is what makes the architecture's *security* claim of §F.1 ("if every tier above an attacker is also captured, the recursion does not save the system") less brittle than a static reading would suggest. The system does not need the right DNA to exist in advance at every tier; it needs *the option for a better DNA to grow into* every tier, and that option is what redemption-driven DNA selection guarantees. The bootstrap problem of §6.5 — Coasys irrigating early DNAs before the holarchy is mature — is therefore not just a monetary bootstrap but an *epistemic* one: enough good-faith DNAs have to exist before the selection mechanism has anything to select among. Once selection is running, it does the corrective work itself.

F. Properties summary

Property	Pure voting	Pure markets	Pure reputation	Natural selection	Pure Shapiro	**Synergy**
Cross-DNA comparison	□	□	□	partial	□	✓ (via scoring)
Selection on the DNA itself	□	□	□	partial	□	✓ (via transitive redemption; cf. §E.1)
Game-proof	□	□	□	partial	□	✓† (proper scoring; DNA-dependent)
Convergence target	majority	Nash	trust loops	local fitness	Nash	≤ Nash (ensemble EFE; below Nash for cooperative-game classes)
Requires common DNA	✓	✓	partial	✓	□	□
Has generative model	□	□	□	□	□	✓
Sybil resistance	partial	✓ (capital)	partial	✓	✓ (conductance)	✓ (conductance + scoring)

Property	Pure voting	Pure markets	Pure reputa-tion	Natural selection	Pure Shapiro	**Synergy**
Truth-telling rewarded	□	□	□	partial	□	✓ (calibrated via stake-volume; cf. §B.1.1)
Backed by	counts	speculation	trust	survival	issuer's goods	scored predictive accuracy + social-graph conductance

†The game-proof claim is *per-flow* incentive-compatible under strict propriety, but is qualified by the open multi-round / collusion surface explored in the next subsection — and, as that subsection makes precise, depends substantively on the *specific Social DNA in use*. The "partial" entries in the natural-selection column reflect that evolutionary dynamics on heterogeneous-task ecosystems give scored truth-telling and cross-DNA comparison only when fitness is *itself* an effective scoring rule — a strong, often unwarranted, assumption. (For reference: Nowak (2006), *Five rules for the evolution of cooperation*, gives a more disciplined treatment of when natural selection *does* deliver each property.)

Qualitatively: aggregating probabilities under a proper rule strictly dominates aggregating binary preferences (voting discards calibration); Shapiro redemption synthesizes liquidity from trust where pure markets require pre-existing liquidity; a scored prediction record is falsifiable in a way that a single reputation score is not; proper scoring permits arbitrarily many local fitness functions to coexist as long as each is locally proper, and the ensemble free-energy reading provides the cross-DNA comparator.

F.1 Game-proof: scope and the open game-theoretic surface

Strict propriety — the property used to mark Synergy ✓ for "game-proof" in the table — guarantees that under a *fixed, single-round* scoring procedure with a *single forecaster* truthful reporting is uniquely optimal. The proposed mechanism is, however, a *multi-round*, *multi-agent*, *redemption-coupled* dynamic, and four manipulation surfaces lie outside what strict propriety alone covers:

1. **Wash-trading** inside an organ — two cooperating accounts repeatedly co-signing each other's flows to mint tokens against fabricated co-signed evidence.
2. **Cartel formation** in YES/NO pairings — if YES and NO sides can negotiate, they can split the pot without either side having an actual claim on truth.
3. **Score-decay arbitrage** — if scores are decayed over time (and the scheme is currently silent on this), agents can game the decay schedule.
4. **Redemption-cascade exploitation** — the §6 walkthrough notes that an Aggregate-NO holder *inherits* sub-organ shares of the aggregate's treasury; a cartel can in principle use this to dump tokens on hostile parties.

The honest statement of the table claim is therefore: *incentive-compatible per-flow under strict propriety, with multi-round and collusion vulnerabilities that depend substantively on the specific Social DNA in use.* Whether each surface is in practice exploitable is a function of the DNA, not of the architecture as such.

For the first three surfaces — wash-trading, cartels, score-decay — the architectural counter-pressure is **fractal stacking plus proof-of-person**. None of these attacks are free in Synergy: each requires sybils (otherwise there is no separate "cooperating account" to wash-trade with) or coordinated insiders. Sybil-resistance (§A) raises the cost of the first; proof-of-person Social DNAs at lower tiers raise the cost of all three by gating who can be a flow-participant in the first place. Higher tiers in the holarchy then run *their own* scoring rules over lower-tier outcomes; if a lower-tier organ's DNA tolerates a wash-trading attack, that organ's outputs will score badly at the parent level, and the parent organ's own redemption dynamics will discount the captured tokens. This is the self-correction story: the architecture is engineered so that a *bad* Social DNA does not propagate — it gets penalized at the next level up by a *better* DNA, and the §E.1 selection-on-DNAs argument generalizes this from a defensive posture against attack to a positive selection pressure that holds whether or not anyone is actively attacking. The catch is that the mechanism rests on the *existence*, somewhere in the system, of a sufficiently well-designed DNA at *some* tier; if every tier above an attacker is captured *at the moment of attack*, the recursion does not save the system in real time. The mitigation is therefore conditional on enough good-faith network growth at the upper tiers — which is exactly why proof-of-person is one of the very first DNAs the holarchy needs to irrigate — but the longer-run answer is that DNA-level selection (§E.1) lets a better DNA grow into any captured tier, so a captured upper layer is not a permanent loss but a head-start for whatever heterodox DNA can demonstrate better calibration.

The fourth surface — redemption-cascade exploitation — is more interesting. The exploit, as the reviewer states it, is that an Aggregate-NO holder inherits sub-organ YES shares "for free" when the aggregate fails, and a cartel could engineer this to dump those shares on hostile parties. The architectural answer is that getting Aggregate-NO tokens *also* requires going through a flow — there is no other way to mint them — and that all currencies in the holarchy are gated through flows in this way. So the cartel has to *do the integration work* to get the upper-level coin in the first place, and that work is itself scored. **If the cartel can in fact reliably get top-level currencies and then redeem them down the cascade for free, that is informationally a signal that the upper-level DNA was wrong** — that whatever flow the upper tier accepted as evidence of competence was gameable, and the downward-redemption is just the system's automatic correction making the right people pay. In other words: the redemption cascade does not have to be defended as a vulnerability; it functions as the calibration signal for the *next* level up. A flow that is followed by predictable mass redemption-cascade extraction *is* the evidence the parent of the parent organ uses to discount that DNA.

We have stated this charitably; it is also, honestly, where the design is least proven. Whether real cartels in real social settings will obey this corrective dynamic, or whether they will route around it faster than the next-level scoring can catch up, is an empirical question that simulation work (cf. §G Q4) is in a position to test. For now: we hold that the redemption cascade is a feature *under sufficient holarchy maturity* and a vulnerability *during bootstrap*. The §6.5 bootstrap discussion of Synergy Fuel as joker — and the irrigation strategy that justifies Coasys minting it — is therefore not just a monetary argument; it is also part of the security argument, since the joker is what lets the holarchy grow upward fast enough to outrun cartel formation at lower tiers.

G. Open questions

The design has four open questions that gate the next version (v0.11). We list them honestly because the article is meant to invite criticism, not to oversell.

Q1 — The SoA slot catalog, and the topology of SoA trees. Two related strands.

The catalog. What are the standard, reusable SoA schema slot types that organs can mount? We have working types (binary / categorical / continuous), but the catalog of *commonly useful claim shapes* — calibration claims, attribution claims, counterfactual claims, normative claims — has not been standardized. This is concrete design work, not theory work.

The topology. Slots in different organs can refer to the same world-fact: a `TuringProbe-Alice` slot in one organ and an `IsPersonhood-Alice` slot in another are both, ultimately, claims about the same Alice. Does that imply a *single global SoA tree* — one DHT into which all organs write into different sub-branches — or *per-organ SoA trees on a shared schema*, with cross-organ co-reference handled by typed identifiers and bridging links? The question matters because it determines what the swap *is* operationally: a write into one shared document, or an atomic two-block transaction across two documents.

The position we take is **per-organ SoA trees on a shared schema, with the swap as a cross-document atomic operation over a blocklace** — *not* a single global DHT. The reasoning is twofold. First, and most importantly, a global tree would force *consensus on slot identity at write time* — and that is precisely the disagreement the cross-DNA mechanism is designed to *defer* until a parent organ can score it. Per-organ trees with shared schema let two organs maintain incompatible bindings of the same world-fact; the parent then learns which DNA was better calibrated by observing how each organ's bindings score against reality. Squashing the disagreement at the storage layer would erase the very signal the system runs on. Second, a single global SoA tree concentrates writes (every organ, every flow, every claim hits one document) and reintroduces exactly the centralizing pressure the architecture is designed to escape — but this engineering objection is the consequence rather than the cause: even if a single global tree could be made to scale, the cross-DNA argument alone is decisive.

The swap, under this position, is a cross-document atomic blocklace operation: it touches one block in the source organ's tree and one block in the destination organ's tree, both signed and causally linked, with no single document needing to mediate. The blocklace's partial order over signed blocks is enough to make the operation atomic; no global ledger is required.

What would have to be true for the global-DHT alternative to be preferable: (a) slot identity globally settleable at write time — which contradicts the cross-DNA premise; (b) writes low-frequency enough that a single ledger does not bottleneck — unlikely at scale; (c) no privacy gradient between local-only and shared content — unacceptable, since local SoA trees are personal-second-brain material that must not auto-share. None of these hold under our intended use; we therefore default to per-organ trees with cross-organ atomic swaps.

Q2 — Generative models per level: how much explicit inference does v0.11 actually run? §C now commits to a position: every level *has* a generative model in principle, and every level approximates it with the local token-holding sufficient statistics in practice. The residual open question is *how much explicit inference* a v0.11 implementation should run on top of those statistics — that is, how much of the hierarchical active-inference message-passing should actually be computed, versus relying on the redemption-driven dynamics to converge the holdings to the right place by themselves.

We see three regimes worth distinguishing:

- (a) *Ledger-only.* No explicit inference. Holdings move only via swaps and redemptions; predictions for the next integration event are read off the holdings directly. **Minimum viable.**

v0.11 ships at least this.

- (*b*) *Light inference.* One or two passes of variational message-passing on top of the holdings, used to issue *forward* predictions for upcoming integration events without changing the ledger. Cheap; possibly worth it for sharpening rare-event predictions.
- (*c*) *Full hierarchical.* Research-grade, costly, probably needs domain-specific approximation. Long-term direction; not committed to.

The minimal commitment for v0.11 is (a). Whether (b) buys enough predictive sharpness to be worth its compute cost is an empirical question we have not answered. (c) is a research direction whose cost-benefit at our scale is wide open.

Q3 — Score aggregation at the organ level. When member predictions disagree on an organ-level integration event, what is the organ's score? Simple sum, precision-weighted sum, best-of, worst-of, leave-one-out CRPS? Different aggregators imply different organ "personalities".

Q4 — Empirical validation. James is running agent-based simulations *without* the Shapiro/scoring layer, to check whether the failure modes we predict actually manifest. The empirical question — does omitting proper scoring really lead to monoculture collapse on heterogeneous tasks? — is settable by simulation.

References

Acemoglu, Daron, and Asuman Ozdaglar. 2011. "Opinion Dynamics and Learning in Social Networks." *Dynamic Games and Applications* 1 (1): 3–49.

Aumann, Robert J. 1974. "Subjectivity and Correlation in Randomized Strategies." *Journal of Mathematical Economics* 1 (1): 67–96.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Buterin, Vitalik, Zoë Hitzig, and E. Glen Weyl. 2019. "A Flexible Design for Funding Public Goods." *Management Science* 65 (11): 5171–87.

Castro, Miguel, and Barbara Liskov. 1999. "Practical Byzantine Fault Tolerance." In *Proceedings of the Third Symposium on Operating Systems Design and Implementation* (*OSDI '99*), 173–86.

DeGroot, Morris H. 1974. "Reaching a Consensus." *Journal of the American Statistical Association* 69 (345): 118–21.

Friston, Karl. 2008. "Hierarchical Models in the Brain." *PLoS Computational Biology* 4 (11): e1000211.

———. 2010. "The Free-Energy Principle: A Unified Brain Theory?" *Nature Reviews Neuroscience* 11 (2): 127–38.

Gneiting, Tilmann, and Adrian E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–78.

Good, I. J. 1952. "Rational Decisions." *Journal of the Royal Statistical Society B* 14 (1): 107–14.

Hanson, Robin. 2003. "Combinatorial Information Market Design." *Information Systems Frontiers* 5 (1): 107–19.

———. 2007. "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation." *Journal of Prediction Markets* 1 (1): 3–15.

———. 2013. "Shall We Vote on Values, but Bet on Beliefs?" *Journal of Political Philosophy* 21 (2): 151–78.

Harris-Brown, Eric, Nicolas Luck, and Arthur Brock. 2018. "Holochain: Scalable Agent-Centric Distributed Computing." Holochain Foundation.

Harsanyi, John C., and Reinhard Selten. 1988. *A General Theory of Equilibrium Selection in Games*. MIT Press.

Jadbabaie, Ali, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. 2012. "Non-Bayesian Social Learning." *Games and Economic Behavior* 76 (1): 210–25.

Kiyotaki, Nobuhiro, and Randall Wright. 1993. "A Search-Theoretic Approach to Monetary Economics." *American Economic Review* 83 (1): 63–77.

Lamport, Leslie, Robert Shostak, and Marshall Pease. 1982. "The Byzantine Generals Problem." *ACM Transactions on Programming Languages and Systems* 4 (3): 382–401.

Lietaer, Bernard. 2001. *The Future of Money: Beyond Greed and Scarcity*. London: Random House.

Lietaer, Bernard, Christian Arnsperger, Sally Goerner, and Stefan Brunnhuber. 2012. *Money and Sustainability: The Missing Link*. Triarchy Press (with the Club of Rome).

Nowak, Martin A. 2006. "Five Rules for the Evolution of Cooperation." *Science* 314 (5805): 1560–63.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.

Pezzulo, Giovanni, Francesco Rigoli, and Karl Friston. 2018. "Hierarchical Active Inference: A Theory of Motivated Control." *Trends in Cognitive Sciences* 22 (4): 294–306.

Ramstead, Maxwell J. D., Paul B. Badcock, and Karl Friston. 2018. "Answering Schrödinger's Question: A Free-Energy Formulation." *Physics of Life Reviews* 24: 1–16.

Ruiz-Serra, Jaime, Patrick Sweeney, and Michael S. Harré. 2024. "Factorised Active Inference for Ensembles of Agents."

Savage, Leonard J. 1971. "Elicitation of Personal Probabilities and Utilities." *Journal of the American Statistical Association* 66 (336): 783–801.

Shapiro, Ehud. 2022a. "A Grassroots Democratic Metaverse." *arXiv Preprint*.

———. 2022b. "Foundations for Grassroots Digital Economies." *arXiv Preprint*.

———. 2023a. "Grassroots Social Networking." *arXiv Preprint*.

———. 2023b. "Grassroots Systems: Concept, Examples, Implementation, Applications." *arXiv Preprint*.

Weyl, E. Glen, Puja Ohlhaver, and Vitalik Buterin. 2022. "Decentralized Society: Finding Web3's Soul." SSRN preprint 4105763.