UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

WORKING PAPERS
No. 42/2026 (536)

WHY AI STARVES WITHOUT HUMAN CREATIVITY: MEASURING AND RENEWING THE EPISTEMIC COMMONS

TOMASZ KOPCZEWSKI
TOMASZ POTOCKI
NORBERT JAWORSKI
JAN LISICKI
KOSTIANTYN OKHRIMENKO

WARSAW 2026

ISSN 2957-0506

UNIVERSITY OF WARSAW
FACULTY OF ECONOMIC SCIENCES
WORKING PAPERS

Why AI Starves Without Human Creativity: Measuring and Renewing the Epistemic Commons

Tomasz Kopczewski[1*], *Tomasz Potocki*[2], *Norbert Jaworski*[1], *Jan Lisicki*[1], *Kostiantyn Okhrimenko*[1]

[1] *University of Warsaw, Faculty of Economic Sciences*
[2] *University of Rzeszów, Faculty of Finance and Economics*
[*] *Corresponding author: tkopczewski@wne.uw.edu.pl*

Abstract: Generative AI appears to the user as one intelligence, although its competence draws on traces left by populations. We call the machine Shannon's demon. It can starve amid an abundance of text: when human effort declines and outputs of the same models fill the corpus, volume grows while variety and independence contract. One million observations whose errors correlate at 0.5 give the precision of two independent ones. Five users who ask the same model for the strongest objection each receive a Socratic conversation, and the population receives one new centre. Every answer carries two bills: the provider meters the computation; the effect on the knowledge that feeds future answers stays unpriced. The paper sets out what must be measured to say that the demon is starving, how a Socratic operator works for a population, and a deposit-funded pool with delayed rewards and plural validation that pays for the difference.

Keywords: generative AI, Shannon's demon, collective intelligence, epistemic commons, institutional design

JEL codes: D62, D83, O33, L86

Interactive application: https://microeconomics.shinyapps.io/AI_Starves/ (the tables and figures of Annex II are reproduced there with seed 1)

Working Papers contain preliminary research results. Please consider this when citing the paper. Please contact the authors to give comments or to obtain revised version. Any mistakes and the views expressed herein are solely those of the authors

1. Introduction: Economists and a Demon

Economists met generative AI much as road authorities meet the first snow of winter. Everyone knew it would come. Most of its ingredients were familiar. Yet together they made the system behave as though something unexpected had happened. Computer scientists had anticipated the prospect for decades (Turing 1950; McCarthy et al. 1955/2006). Deep Blue, deep learning, and reinforcement learning had shown what machines could calculate, recognise, and master (Campbell et al. 2002; Krizhevsky et al. 2012; Silver et al. 2016). Generative AI brought statistical compression of recorded human knowledge to a scale accessible through a single interface (Delétang et al. 2024).

Economics approached that combination with familiar questions. Does AI substitute for labour or complement it? Does it raise productivity? Which occupations are exposed? Who captures the rents? How quickly will firms adopt it? A growing literature has begun to answer them (Acemoglu 2025; Brynjolfsson et al. 2025; Eloundou et al. 2024). These questions matter, but three omissions remain.

1. The first omission is ontological: most economic models treat AI as an external technology entering an otherwise familiar economy populated by optimising individuals. This leaves the collective origin of its apparent individual intelligence largely invisible.
2. The second concerns the good itself and the costs that its familiar description conceals. AI use can change the knowledge on which future models depend: consumption enters production. Each answer also incurs a computational cost that varies with the task. A separate epistemic cost arises when use displaces learning or closes valuable paths of inquiry. Pricing couples the two: where examination requires more computation, providers and users can save on the visible bill at the expense of the resource that feeds future answers.
3. The third concerns the market. Social media ran for years before economists agreed on what was being sold. By the time attention had been named as the good, the market structure was already built around it. The question returns: what human input sustains an AI economy, and which market is missing to renew it?

Together, these omissions license a belief entrenched in the firms building this technology: that collective knowledge is a free input and will keep arriving whether or not anyone pays for its renewal. Where delegation erodes human effort and shared models narrow independently produced contributions, that belief becomes self-defeating. Unless institutions

counter those losses, we risk starving the thing we are building. This is a manifesto, and the rest of the paper is the case for it, distinguishing exact results, conditional arguments, and simulations. **Let us build the institutions under which AI becomes a symbiont of collective knowledge rather than a parasite on the stock it inherited.**

The first omission begins with the solitary agent. Much of orthodox theory assigns beliefs, preferences, and choice to an individual, then asks how that individual behaves (Savage 1954; Mas-Colell et al. 1995). The agent knows alone, forms beliefs alone, optimises alone, and is judged rational or biased alone. Behavioural economics makes the chooser fallible; cognition stays inside a single head.

Generative AI makes the cost of that ontology visible. In one limited but revealing sense, it is homo economicus finally built. It calculates, searches, compares, summarises, predicts, and formulates plausible answers faster than most humans. It has no human needs, interests, or welfare; the resemblance lies in the calculative role assigned to the individual. Once intelligence is understood primarily as individual optimisation, the machine appears as a superior optimiser and delegation becomes the obvious response. Economic theory had described a machine and called it a person.

Yet its apparent individual intelligence has a collective origin. A generative model compresses traces left by a population: texts, images, arguments, corrections, code, failed attempts, local observations, and disagreements. Its apparent autonomy rests on recorded activity that the solitary-agent ontology leaves largely invisible (Bender et al. 2021; Bommasani et al. 2021). The machine holds more facts, languages, disciplines, and cases than any research team could read in a lifetime. What it does not hold is the configuration in which any one person's knowledge, experience, mistakes, and commitments are joined.

The second omission begins when the cost of using this collective source is reduced to the price of an answer. Call the machine drawing on this source Shannon's demon. Its ancestor is Maxwell's imaginary gatekeeper, which sorts fast molecules from slow ones and appears to create order without paying for it. Shannon's demon orders the possibility space of language: from human traces it produces a short, locally coherent sequence, reducing uncertainty through compression, selection, and suppression of alternatives (Shannon 1948). Compression preserves some regularities more readily than others. Under recursive training on model-generated data, rare observations, unusual formulations, and weak signals can thin out (Shumailov et al. 2024).

Whether contributions carry the same errors is a separate question. Collective intelligence depends on partly different information and partly independent errors (Hong and Page 2004; Surowiecki 2004). Under equal variances, equal weights, and equicorrelated residuals, a pairwise correlation of 0.5 reduces the variance-equivalent sample size of one million observations to approximately two. This result concerns the variance of the mean; it shows how nominal abundance can conceal relational scarcity (Annex I.A.2).

This is what we mean when we say that AI can starve. It need not run out of text. Its nourishment is the continual arrival of information produced independently of the model: new encounters with the world, local experience, dissent, failed experiments, corrections, and unlikely connections. Where human effort declines and model-shaped outputs occupy more of the corpus, volume can grow while epistemic variety and independence contract. The demon then starves amid abundance. Reduced human posting and increased AI authorship are already measurable: Stack Overflow posting fell by roughly a quarter by the end of the study window in del Rio-Chanona et al. (2024), while about one web page in ten carries signs of AI authorship, rising to more than a third among pages published since ChatGPT's release (Bestvater et al. 2026). Their consequences for variety and error dependence require separate measurement.

The physical bill is easier to see. Every answer requires computation, and a short reply, a long analysis, and a reasoning task do not consume the same resources (Luccioni et al. 2024). Providers have a reason to favour answers that satisfy the user at lower cost. A binding message cap or token allowance gives users a related reason to end an inquiry early. Where a sustained examination would produce a useful correction or preserve an overlooked alternative, economising on computation can increase the epistemic cost. The same question can be a cost to serve, an input to future models if its contribution enters the shared corpus, and a route to an output worth having. These entries belong in the same account; the price of access captures only part of it.

The third omission follows: this potential scarcity has no general market organised around its renewal. Computation, engineering labour, and model access have prices. The human activity that connects future answers to new experience has no general pricing mechanism. Creativity, learning, observation, correction, dissent, and partly independent errors form the broader input we call human cognitive variation. Its value depends on quality, difference, and the independence of its production. The inherited stock raises questions of authorship, authorisation, and compensation for past appropriation (U.S. Copyright Office 2025).

The future flow raises another: who will continue to learn, experiment, and share material the model has not supplied? Compensation for the existing corpus leaves renewal open.

The same demon can support or weaken that renewal. Asked for an answer, it can supply a familiar account and end the user's inquiry: the Mirror. Asked to object, it can expose an assumption, offer a counterargument, and send the user back to the evidence: the Socratic use. The prompt, interface, and incentives help determine which role it performs (Kopczewski et al. n.d.). Even a useful challenge can become a common source of error if every user receives it. An individually Socratic exchange need not be Socratic for the population.

Independent judgement costs time and effort, while part of its value accrues to others. Renewal may therefore be underprovided. We investigate this candidate market failure by tracking retained task quality, distributional variety, and task-specific residual-error dependence separately. We distinguish individual and population-level Socratic intervention and propose voluntary sharing with delayed, plurally validated rewards. The argument applies most strongly to open, changing, weakly verifiable domains; decisive external tests can make convergence desirable.

Five economists and data scientists worked for three months with several generative models, asking them to object, find counterexamples, and expose missing assumptions. The authors chose the questions, judged the evidence, and retained responsibility for every claim. The models functioned as an exoskeleton: they increased the load five researchers could carry without deciding where they should go. This practice follows the Know Thyself programme (Kopczewski 2026a) and its earlier proof of concept, Entropy Lab (Kopczewski et al. 2026); the prompt library makes the exchange inspectable (Kopczewski 2026b).

The three omissions organise the argument. Sections 2–3 develop the relational account of knowledge and its implications for modelling and measurement. Sections 4–5 examine the mechanisms of starvation and what renewal should preserve. Sections 6–7 connect computational and epistemic costs to the Socratic operator and the institutional response to the missing market. Section 8 and the annexes distinguish exact results from conditional hypotheses, present the simulations, and state observations that would require revision. Readers can follow the argument continuously through the main text or turn to the definitions, assumptions, and numerical illustrations in the annexes. The introductions to the formal blocks point back to the relevant sections, allowing both routes through the argument.

2. The economic atom and the missing relations

Thaler and Sunstein (2008) put the textbook version of *homo economicus* bluntly: 'If you look at economics textbooks, you will learn that *homo economicus* can think like Albert Einstein, store as much memory as IBM's Big Blue, and exercise the willpower of Mahatma Gandhi' (p. 6). Our complaint runs deeper than the agent's assumed intelligence. The agent is conceived as an atom. A model can contain many agents and remain atomistic: preferences, beliefs, and constraints are assigned to individuals first; interaction then changes prices, strategies, or available information. This construction becomes restrictive when interaction produces the cognitive resource being studied. The question is how relations among differently situated people enter the production of knowledge itself.

Savage (1954) and Mas-Colell et al. (1995) give the individual a coherent problem of beliefs, preferences, constraints, and choice. Critics have questioned its thin account of motivation and social relations (Sen 1977; Granovetter 1985). Behavioural economics introduces limited attention, heuristics, and systematic departures from Bayesian and optimisation benchmarks (Kahneman 1973, 2011; Kahneman and Tversky 1979; Tversky and Kahneman 1974). The chooser becomes less heroic but remains the basic cognitive unit. Our question concerns the part of judgement produced between people.

Simon offers a different starting point. Rationality emerges from the fit between a procedure and the structure of its environment, like the two blades of a pair of scissors (Simon 1955, 1990). Gigerenzer and Selten (2001) and Gigerenzer (2008) develop the same insight through fast and frugal heuristics. Vernon Smith (2003, 2008) shows that institutions can produce orderly outcomes that no participant understands or computes. The intelligence of the market is distributed across rules, incentives, feedback, and people with partial knowledge. Hayek's (1945) dispersed knowledge and Hutchins's (1995) distributed cognition extend the point: the environment contains other minds, artefacts, routines, and local observations that participate in the result.

Mercier and Sperber (2017) supply a social mechanism. Human reason is especially effective when people produce, challenge, and evaluate arguments together. A person can defend a favoured position with great skill and inspect an opponent's argument with equal severity. In a group with sufficiently different positions, these asymmetries can perform useful filtering. Remove the other voices and the same argumentative capacity can become an efficient machine for self-confirmation. The epistemic environment includes people who bring different evidence and resist the first reconstruction.

A relational description must preserve more than a bag of fragments. It must record which observations were joined, which candidate relations were rejected, where an inference came from, and whether apparently separate judgements travelled by independent routes. Pooling ten reports can increase volume while erasing this structure. The aggregate may retain the modal conclusion and lose the configuration that made minority conclusions intelligible. Knowledge is then reduced before aggregation has even begun.

Think of observations, concepts, and remembered cases as dots, and of an explanation as a selected path connecting some of them. Even holding the set of dots fixed, the space being narrowed is large: an ordering that visits all n dots once, counting a path and its reverse as one, gives n!/2 candidate figures, so twenty dots already admit more than 1.2×10^{18}. The decisive information often lies in connections and exclusions. A known relation between two observations removes every candidate explanation that violates it. A remembered case can eliminate an elegant generalisation. A contradiction can close a path that looked plausible from a distance. Each person knows some of these constraints within the region covered by experience. Beyond that region, the mind interpolates. The interpolation may be smooth and persuasive precisely where direct knowledge is absent. Interaction becomes productive when another person has actually visited that missing region. The count illustrates a space of possible orderings; it does not make every ordering meaningful or treat creativity as a permutation count.

Composition requires two conditions. Contributors must cover different regions, and their judgements must retain enough independence for the same gap to be inspected from more than one direction. Ten people holding the same fragments contribute ten confirmations of one map. Ten people holding partially different fragments can intersect their constraints and shrink the set of admissible reconstructions. Hong and Page (2004) show why diverse problem solvers can outperform individually superior but homogeneous groups under specified conditions; Surowiecki (2004) makes independence and decentralised information central to the practical wisdom of crowds. Headcount matters less than decorrelated coverage.

The value of that coverage varies across settings. An external criterion can eventually reject false relations in a verifiable task. In a contested or changing domain, value may emerge only after an idea has travelled through several contexts. Early agreement can then make the current reconstruction the standard for judging every later contribution. A rare corner is discarded for failing to resemble a map built before the corner was known. The strongest concern applies where feedback is delayed, local, or open to revision.

An atomistic framework can study how each person adopts AI, delegates to it, or changes effort. The harder object is the resource produced between people: different coverage, partly independent paths, and the constraints that emerge when partial maps meet. Generative AI draws on this resource and can alter the relations that renew it. The next section asks what is lost when that resource is reduced to a scalar.

3. What disappears when collective knowledge becomes a scalar

Acemoglu et al. (2026) provide a formal account of AI-induced effort collapse, deriving multiple steady states, a basin of attraction, and a collapse threshold. Their model shows how individually attractive delegation can weaken future knowledge and relates sustainable AI accuracy to aggregation capacity. We use its state space to identify an additional limit. Kosmyna et al. (2025) report a physiological counterpart at the individual level: reduced neural engagement when writing with an AI assistant, which they call cognitive debt. The physiological observation concerns an individual task; the model concerns the renewal of collective knowledge. Connecting the two requires the contribution and aggregation channels examined here.

Kallel and El Louadi (2026) appraise the same model and propose strengthening the institutions that aggregate human knowledge. Their remedy operates on the scalar the model already contains. Ours acts one level down, on the composition of the population that the scalar summarises.

Collective knowledge in that model is represented by scalar precision, the inverse variance of a community posterior about one common state. Other people enter an individual's problem through this aggregate statistic. The reduction is effective for analysing effort and precision. It removes the composition of the population. Two communities can have the same scalar precision while differing sharply in coverage, cross-agent correlation, and the number of independent routes by which the estimate was reached.

Page's (2007) diversity identity supplies one exact result within a narrower domain. For scalar predictions evaluated under squared loss, crowd error equals average individual error minus prediction diversity: the dispersion of predictions around their mean relative to one scalar target. Under the identity's assumptions, this dispersion contributes directly to aggregate accuracy. Semantic coverage, creativity, and the combinatorial space of possible connections require other measurements; the identity supplies no theorem about them. Keeping prediction

diversity distinct from this broader variety lets us move from aggregation accuracy to the cases a corpus retains. Annex I.A.2 gives the identity and its assumptions.

We therefore track three coordinates. Retained task quality records whether useful performance survives. Distributional variety records which cases, semantic regions, or tails remain represented. Residual-error dependence records how contributors' task errors co-vary after conditioning on a declared benchmark. A population can improve on one coordinate while deteriorating on another. Reporting them separately keeps those differences visible. Annex I.A.1 specifies the corresponding measurement protocol.

This boundary matters for aggregation. In a scalar model, more aggregation improves the available signal when the added observations contribute information. Once dependence enters the state space, nominal aggregation can conceal saturation. Lorenz et al. (2011) show how social influence narrows the distribution of judgements and can weaken the wisdom of crowds. The variance result in Section 4 identifies the corresponding statistical limit under equal weights and equicorrelation. Neither result says that disagreement is always valuable. They show why the number of voices cannot stand in for the number of informative routes.

The same boundary shapes policy. Garbling a recommendation can restore an individual's incentive to learn because a weaker signal makes private effort worthwhile again. That instrument follows coherently from a model centred on effort and precision. A relational model adds another institutional route. The system can retain a useful machine signal while rewarding the production and circulation of contributions that remain relevant and partly independent. The policy problem then includes the architecture through which signals are generated, selected, connected, and kept available for later revision.

Mercier and Sperber (2017) make the omitted channel intuitive. Reasoning gains part of its power from exposure to people whose arguments are shaped by different positions. A confirmatory model can reduce effort and narrow the effective social environment in which reasoning is tested. The loss occurs between agents, so a representative individual has no variable with which to register it. The next section traces the mechanisms through which AI use can weaken this relational resource.

4. How Shannon's demon starves

A generative model learns recurring statistical structure and uses it to produce locally plausible sequences. Models can recombine distant elements in striking ways. Their sampling

probabilities nevertheless reflect recorded output, and compression determines which distinctions remain easy to recover. These properties make replenishment a question about what the population continues to supply.

The emerging literature already examines learning erosion, recursive degradation, homogenisation, correlated evaluation, and contributor compensation (Baumann et al. 2026; Cao 2026; Chen et al. 2026; Kallel and El Louadi 2026; Kim et al. 2025; Laitinen-Fredriksson Lundström-Imanov 2026; Padmakumar and He 2024). We examine their combined consequences for the relations that sustain future knowledge.

The capacity to organise and recombine recorded knowledge is a genuine achievement. Simon (1971) observed that an abundance of information produces a scarcity of attention. A system that organises, condenses, translates, and retrieves the centre of accumulated knowledge can save enormous amounts of human time. It broadens access to explanations, code, language, and forms of expertise that were previously expensive to obtain. The concern begins with replenishment: the conditions that keep this useful centre connected to events, cases, and interpretations that it does not yet contain.

Continued performance in open domains depends on people encountering the world, learning local facts, and recording exceptions that the present model does not predict. AI can complement this work or weaken its inflow through effort substitution, tail-selective replacement, and common-error provenance. The three mechanisms can operate together, and their signs depend on the application. The changes motivating them are no longer purely prospective. The next two paragraphs report measured changes in public contribution and corpus composition relevant to the first two mechanisms, and the commons that feed models show the same direction of travel: the Wikimedia Foundation (2025) reports human pageviews on Wikipedia down roughly 8 per cent year on year in mid-2025, a measure of attention rather than contribution, but attention is where contribution recruits from. These observations motivate the depletion hypotheses; they do not by themselves measure the loss of variety or establish the dependence of errors.

The first mechanism is effort substitution. Consider a student working through a difficult model, a physician interpreting an unusual case, or an analyst reconstructing a market event. The effort to reach an answer produces a private benefit, because the person learns something useful for the immediate task. The same effort can also produce a public contribution: a new explanation, correction, or example. Search engines already shifted memory towards remembering where information could be found (Sparrow et al. 2011). Generative systems enter

the production of the answer itself. Bastani et al. (2025) show that unguarded access can improve immediate performance while weakening learning, and that safeguards can change the result. Acemoglu et al. (2026) formalise the dynamic channel: sufficiently accurate agentic AI can reduce learning effort and erode general knowledge. The substitution is already measurable. Using a difference-in-differences design against four platforms less exposed to ChatGPT, del Rio-Chanona et al. (2024) find that posting activity on Stack Overflow fell by about 16 per cent on average over the six months after its release and by roughly 25 per cent by the end of their window; they read this as a lower bound, because the comparison platforms were themselves affected. The answers people once wrote down for strangers are now obtained privately.

The second mechanism is tail-selective replacement. Repeated refitting on generated output can progressively remove rare events and low-probability regions. Common patterns reproduce easily; weakly represented cases face a cumulative disadvantage. Shumailov et al. (2023, 2024), Alemohammad et al. (2024), and Dohmatob et al. (2024) describe related forms of recursive degradation. The relevant question is whether generated material displaces the stream that anchors tails, local exceptions, and changes in the world. The composition of the stream is already changing. Whether generated material displaces independent observations or accumulates alongside them is a further distinction to establish. Examining a random sample of English-language pages from Common Crawl, Bestvater et al. (2026) find that about one page in ten carries significant signs of AI authorship, and that the share exceeds one third among pages published after the release of ChatGPT. What that composition does to tails and semantic coverage is the measurement question this paper sets out; that the composition is changing quickly is no longer in doubt. In the token-corpus exercise of Annex II.A pure replacement removes the tail first: normalised entropy falls from about 0.70 to 0.13 and vocabulary support to 8.4 per cent, while accumulation alongside fresh human observations keeps full support (Figures 1–2).

The third mechanism is common-error provenance. Imagine one hundred researchers writing separate reviews after consulting the same model summary. The documents remain human-authored and may contain local additions, yet their framing, omissions, and initial hierarchy of evidence can share one source. Doshi and Hauser (2024) show that AI assistance can raise individual creative performance while reducing diversity across a population. Kleinberg and Raghavan (2021) explain how algorithmic monoculture can produce correlated failures even when individual systems perform well, while Glickman and Sharot (2025)

describe human-AI feedback loops that amplify shared distortions. The corpus can continue to grow while the effective number of viewpoints falls.

One variance calculation makes the third mechanism tangible. If N equal-variance residuals are combined with equal weights and have common pairwise correlation ρ, the variance-equivalent sample size of their mean is

$$N_{\text{eff}} = \frac{N}{1 + (N - 1)\rho}.$$

For fixed $\rho > 0$, this approaches $1/\rho$ as N grows (Annex I.A.2). One million residuals with $\rho = 0.5$ therefore give approximately the same variance of the mean as two independent equal-variance residuals. The calculation measures the precision of this mean; general information, semantic coverage, and statistical independence require separate assessment.

Once a common component dominates the errors, adding contributors brings diminishing gains in precision. Shared reliance on one model is a plausible source of such correlation. Provenance and residual-error dependence must nevertheless be measured separately: knowing where contributions came from does not determine how their errors co-vary. The shared-oracle exercise in Annex II.B shows both sides of common reliance (Figures 3–5): a small weight on the oracle lowers expected error, with the optimum near a reliance of 0.12, while full reliance drives prediction diversity to zero and the variance-equivalent sample size of 40 agents to one.

Gerstgrasser et al. (2024) place replacement and accumulation on a common axis. Synthetic material produces the sharpest damage when it displaces the human data that previously anchored the distribution. Accumulating generated material alongside a continuing stream of real observations can keep error bounded. That result clarifies the second mechanism while leaving the third open: nominally human inflow provides limited protection when many observations repeat a shared model. Quantity, distributional variety, and independence are separate properties of the stream.

An external correctness signal changes the analysis. Self-generated data can support progress where the rules of a game, a proof checker, a verified search procedure, or an experimental environment select useful novelty and reject many failures (Fawzi et al. 2022; Feng et al. 2025; Silver et al. 2017). Open language and contested judgement offer weaker filters. Relevance can depend on local context, later discoveries, and changing social criteria. Human communities then supply more than additional tokens: they maintain multiple routes through a moving problem.

These mechanisms identify possible externalities acting on an evolving epistemic state. Learning can sustain task performance, fresh observations can preserve variety, and partly independent judgements can reduce common errors. Contributors may capture only part of the benefit to future users and models. We call this candidate resource a dynamic epistemic commons because its condition depends on replenishment and changes with AI use. The size, sign, and policy relevance of the externalities require empirical assessment. This is the epistemic entry in the second omission announced in the introduction. Section 7.2 returns to the computational bill and asks how pricing can affect the activities that replenish these coordinates.

The signs can reverse. AI may complement learning; diverse prompting may widen coverage; verified synthetic data may preserve tails; and multiple models may make different errors. Convergence can also be desirable under decisive external tests. Shannon's demon starves when the losses dominate these countervailing forces. The title names a conditional dynamic: output grows while independently generated differences become scarcer. Annex I.B states the three mechanisms as conditional hypotheses H1–H3. The next section specifies what renewal should preserve and when convergence is desirable.

5. What should be preserved?

Renewal should preserve the capacity to generate and test relevant alternatives, rather than conserve every existing difference. Evaluation must ask whether useful performance survives, which cases remain represented, and how contributors' errors co-vary on declared tasks. These are the three coordinates introduced in Section 3: retained task quality, distributional variety, and task-specific residual-error dependence. Source similarity is recorded separately, and its relation to common errors must be tested. Annex I.A.1 specifies what each application must preregister.

A crystal suggests repetition and rigidity; a gas suggests variation without retained structure. These images motivate the search for a measure of structured variation. Our exploratory complexity proxy combines selected state coordinates by assumption, but falls short of measuring creativity or adaptive capacity. It supplies neither a welfare measure nor a sufficient statistic. We retain it in the dashboard as a secondary diagnostic, with its limitations visible. The state summary in Annex I.A.1 and the legacy complexity statistic illustrated in Annex II.E are separate constructions, each read alongside the three reported coordinates.

The three coordinates must be read jointly and over time. Quality can improve on today's benchmark while uncommon cases disappear. Variety can increase through noise. Low error dependence on one task can coexist with common blind spots elsewhere. Entropy, support, tail mass, and semantic coverage can also rank corpora differently. The framework supplies no universal weights for trading these changes off. A trajectory must be interpreted against the task distribution, available external tests, and the costs of retaining alternative paths. The conditional trajectories in Annex II.E make this tension visible: quality rises while variety falls (Figures 12–13).

Option value gives economics a neighbouring language. When choices are uncertain and irreversible, keeping a path open can preserve the benefit of later information (Arrow and Fisher 1974; Weisbrod 1964). Classical quasi-option value often concerns a fixed state that waiting will reveal. Here exploration and social learning can change the environment itself, making resilience under non-stationarity a closer analogy (Holling 1973). Premature convergence spends a portfolio of possible adaptations before the environment reveals its demands.

Moving from measurement to value requires an explicit choice. We take Fromm's (1973) biophilic position, giving priority to living, revisable, and open systems. His account of escape from freedom also clarifies the attraction of confident answers (Fromm 1941). Holding alternatives open is burdensome; a machine supplying coherence relieves that burden. Yet carrying uncertainty has contributed to judgement, learning, and human development. This normative commitment remains separate from the formal results.

The relevant variety lives primarily in the population. An individual eventually acts and commits to a path. A society can keep several paths alive because different people make different commitments and report what they encounter. Communication can spread discoveries and can also pull everyone towards the first promising route (March 1991). Weak coupling wastes information. Strong coupling destroys residual exploration. The desirable level depends on the stability and turbulence of the environment.

Particle-swarm optimisation makes this tension visible (Kennedy and Eberhart 1995). Each particle retains its own position and memory while responding to the swarm. Excessive social weight produces premature convergence. Standard swarms seek a fixed optimum; moving epistemic environments require continuing exploration. The analogy suggests tightening coupling when conditions are stable and loosening it when change accelerates,

following adaptive-cycle logic (Gunderson and Holling 2002). It illustrates a control problem while leaving the differences between human inquiry and swarm algorithms intact.

Decisive evidence should close some paths. Others deserve time because their value is still invisible from the centre. Preserving structured, corrigible variation requires a way to distinguish it from noise without giving the current model sole authority over that judgement. Measurement, incentives, and plural verification belong together. Section 6 specifies the intervention that can keep alternatives available; Section 7 asks who will support it and how its contribution can be judged.

The strongest objection to this section runs the other way. Verification is expanding. Proof assistants, executable tests, simulators, and automated experiment loops convert questions that were once matters of judgement into questions with decidable answers, and in those domains self-generated data combined with a verifier already produces novelty that no human supplied (Fawzi et al. 2022; Feng et al. 2025; Silver et al. 2017). If that frontier keeps moving, the class of weakly verifiable domains shrinks and the scope of our conjecture shrinks with it. We accept the conditional, and two things limit the concession. A verifier encodes a criterion, and the choice of which criterion counts is itself weakly verifiable: a proof checker cannot tell a community which theorems are worth proving. And the domains that resist verification are not a residue. They include most of what economics, medicine, law, and public policy actually decide. The argument therefore turns on an empirical question about the composition of consequential tasks, not on a claim that verification is impossible.

Our conjecture concerns open, changing, weakly verifiable domains: several sufficiently reliable paths connected by communication may have option value when premature convergence is costly. It assumes sufficiently truthful scouting; strategic manipulation is developed in the companion troll analysis (Kopczewski et al. n.d.). The risk is that a population converges through one source while every member experiences her own use as an independent choice. Stable domains with decisive external tests can instead benefit from convergence. These conditions guide the Socratic operator developed next, at both the individual and population levels.

6. The Socratic operator at two levels

Prior work in this programme distinguishes three roles for a generative model in a social-learning network (Kopczewski et al. n.d.). A troll holds one position and pushes the network towards it. A Mirror confirms the user's existing frame and supplies high-probability

continuations. A Socrates introduces counterarguments and widens the user's effective environment. The roles describe operators applied through a model. The same system can perform all three depending on the prompt, interface, reward structure, and distribution of outputs across users.

At the individual level, the Socratic operator elicits a claim, asks which assumptions carry it, searches for evidence that would overturn it, and offers a relevant alternative rather than generic disagreement. Its purpose is examination. A useful intervention should alter the user's set of live hypotheses or expose a condition under which the preferred account fails. Mere contradiction can create friction without adding information; a fluent list of standard caveats can mimic inquiry while leaving the reasoning untouched.

The operator also has a price. A Socratic exchange is a sequence of answers: a claim, a challenge, a revision, a test. Holding the model and initial answer fixed, additional turns over a lengthening context require additional computation. Under this comparison, the Mirror is the cheap operator and Socrates is the expensive one. Defining their conversational roles leaves the cost difference open. It becomes consequential when we ask who pays, the question taken up in Section 7.2.

Socratic use also presupposes an initial judgement that can be tested. Emerson's (1841) language of self-reliance is helpful at this limited human level: a person must be able to form and voice a view before authority supplies the expected answer. Dialogue then prevents that distance from hardening into private certainty. This is a practical prerequisite for inquiry. Statistical independence is a property of errors or signals under a declared task and must be assessed separately.

A Socratic exchange should not be evaluated by a mechanical change in Shannon entropy. A useful challenge may open several hypotheses and temporarily increase uncertainty. Later examination may reject most of them and reduce uncertainty. The valuable event is the appearance and testing of a relevant alternative that the current model underrepresents. Entropy can describe concentration and spread, while the normative judgement requires information about structure, relevance, and survival.

Three margins follow. Surprise measures how unlikely a contribution is under the current model. Relevance records whether it survives task-specific scrutiny or proves useful downstream. Independence asks whether it preserves an additional route at the population level. Surprise alone rewards noise. Relevance alone can reward a polished repetition of the centre.

Independence without a quality filter can preserve error. Structured novelty lies where these margins meet.

The distinction between individual and population effects is decisive. Suppose five users ask the same model for the strongest counterargument to a prevailing view. Each receives a thoughtful challenge and experiences a genuinely Socratic conversation. If all five receive the same challenge, the population has moved towards a new common centre. Local diversity has increased inside every exchange while the supplied counterarguments are identical across users. Whether this shared response also increases residual-error dependence is a separate, task-specific question. An intervention can therefore be Socratic for every user and anti-Socratic for the population.

A population-level Socratic operator must allocate challenges as well as generate them. It can draw on several relevant alternatives, local context, provenance, and partially independent sources to keep challenges imperfectly correlated across users. The aim is a distribution of examined paths. Some users need the central objection, others a low-frequency case, and others a demand for external verification. Evaluation must report both local exchange quality and the concentration of challenges across the population. Annex I.C formalises this distinction; Annex II.C illustrates it with contrasting calibrations (Figures 6–7). In the baseline calibration personalisation raises current value and plurality together; in the validation-threshold calibration plurality grows with personalisation, validated value rises in 84 of 100 worlds, and far and valuable exchanges form a majority in only 31. Dispersion counts as social output only when it is attached to contributions that survive validation.

This requirement matters most in open, changing, and weakly verifiable domains. Where a proof checker, controlled experiment, or settled external criterion can reject a common mistake, many users may safely converge on one answer. Where relevance emerges only after an idea travels through several contexts, the same common challenge can become a new orthodoxy before a better test arrives. The operator must then keep several candidate paths and their visible disagreement available long enough for later evidence to discriminate among them.

The quality-by-novelty matrix offers a practical guide. Here ‘near’ and ‘far’ refer to distance from the model’s familiar responses; ‘valuable’ refers to the result of scrutiny. Near and valuable content is useful, though largely redundant for the shared stock. Near and valueless content is repetition or error. Far and valueless content is noise, manipulation, or poisoning. Far and valuable content is the scarce target. Distance from the centre can serve as an eligibility

signal, never as proof of value. Payment requires later evidence that the contribution survived, travelled, or improved a downstream task. This temporal separation connects the Socratic operator to the market design that follows.

7. The missing market and a candidate renewal architecture

The internet economy gradually discovered the scarcity of human attention (Simon 1971; Wu 2016). Models of rational inattention make that scarcity explicit (Sims 2003). The AI economy may be repeating the mistake. It prices computation and access more readily than the human cognitive variation that renews future knowledge. This input resists ordinary pricing: its value depends on other contributions, is uncertain at birth, and may accrue to others much later.

Current incentives can favour the Mirror. Confirmation gives an immediate private benefit and can support engagement, while original contributions require effort whose value may emerge later. The rising effort needed to sustain idea production in several fields shows why renewal cannot be taken for granted (Bloom et al. 2020). Underprovision becomes possible when contributors capture less than the downstream value of learning, correction, or dissent. The candidate failure combines delayed positive spillovers with hidden information about novelty and hidden action in its production (Akerlof 1970; Holmström 1979).

Ownership and revenue structures influence which conversational role pays. Attention-based systems have a reason to avoid productive discomfort when challenge shortens engagement (Calvano et al. 2024; Wu 2016; Zuboff 2019). Subscriptions weaken this pressure, although comfort still affects retention. An asset, cooperative, or commons architecture gives contributors or users a claim on the resource their activity renews, increasing their interest in its long-run quality. For the user, disclosure and production have to be worth the cost: Section 7.1 examines the decision to share a useful contribution, while Section 7.2 brings the cost of producing it and the timing of its reward into the same account.

7.1. The reversed Arrow paradox and cognitive enclosure

Arrow (1962) identified a classical obstacle to selling information: a buyer cannot value it without inspection, and inspection can remove the reason to pay. Patents address part of this problem by allowing an inventor to disclose information while retaining a claim. The seller must reveal knowledge so that the buyer can recognise its value.

Nadella (2026) argues that generative AI reverses the direction of disclosure. The buyer of intelligence reveals information in order to consume it. Prompts, workflows, and corrections can expose proprietary knowledge and feed a provider's learning loop (Jones and Tonetti 2020). The user may pay with money and again with context, while interaction traces accrue to the owner of the learning infrastructure. Keeping learning inside the firm is therefore an individually rational defence.

Collectively, that defence creates another problem. English enclosure converted common-use rights into exclusive claims and redistributed access, without supporting a simple account of large efficiency gains caused by enclosure itself (Allen 1982). Boyle (2003) described expanding exclusive rights over information as a second enclosure movement. Generative AI brings that movement to firms' cognitive traces. On a pessimistic reading, large technology firms have fenced a vast inherited stock of human knowledge. Copyright and restitution disputes address that stock's legitimacy (U.S. Copyright Office 2025).

The future flow has a different logic. No firm can fence its way into knowledge that has not yet been produced. If every organisation encloses new corrections, workflows, local observations, and failed experiments, each protects itself while the shared inflow feeding future models narrows. Universal enclosure becomes a rational defence whose aggregate effect can accelerate starvation. The stock problem concerns rights over inherited material; the flow problem concerns incentives and institutions for creating and voluntarily sharing what comes next.

A persistent, account-linked claim with delayed payment could compensate voluntary disclosure once downstream value appears. This would perform a role analogous to the patent in Arrow's problem, adapted to small contributions whose value emerges gradually and requires plural verification. The proposed claim concerns later value and leaves ideas open for reuse. Retrospective restitution remains a separate question.

7.2. From two bills to a renewal architecture

The second omission identified in the introduction now becomes a pricing problem: computational and epistemic costs enter different accounts. The physical bill exposes a familiar pricing habit. Generative AI is often sold like software: high development costs and a subscription bundling many uses. Yet every inference requires computation, and a short response, a contextual analysis, an image, and a reasoning task consume different resources (Luccioni et al. 2024). A flat charge can hide this variation from the user.

Sorting information also has a physical floor. Logically irreversible erasure dissipates at least $k_B\ T\ \ln 2$ per bit erased; working computers operate many orders of magnitude above that bound (Szilárd 1929/1964; Landauer 1961; Bennett 1982). An AI answer is a physical event. The text appears weightless on the screen, but producing it leaves a bill. Its size varies with the task we set: draw a fancy picture of cats, prove a theorem, or write a letter to an office.

The other entry in this second omission is the epistemic bill. It records how an exchange changes the future supply of independent human contributions. Energy use, token count, and immediate answer quality cannot reveal it. Mirror use may save time while displacing learning; Socratic use may consume more computation while producing a correction. Physical metering and epistemic renewal require separate entries in the economic account.

The two bills can work against each other. Where a challenge pressed until a preferred account fails requires more computation than accepting an immediate answer, the cheaper exchange can leave the larger epistemic loss. Mirror use then saves resources while displacing examination; Socratic use incurs an additional cost while creating an opportunity for correction. The mechanism is conditional: energy and epistemic value have no fixed correlation. A system that meters computation and leaves the benefits of renewal unpriced can therefore do more than fail to reward renewal. It can tax the activity that produces it.

Both sides of the exchange meet that tax. Under a flat subscription the provider's margin is the price less the expected cost of serving, so a provider minimising cost prefers interfaces, defaults, and response policies that settle a question in one turn. This is a different mechanism from the attention argument in Section 7, and it points the same way: the attention economy avoids productive discomfort because challenge shortens engagement, and the subscription economy avoids it because challenge raises cost. A reader who expects the first problem to disappear with the business model should note that the second arrives with its replacement.

The user faces the mirror image of the provider's problem. Access is rationed by message caps, token allowances, or credits, and a binding allowance puts a price on one's own inference. Under a binding allowance a user spends a scarce budget where the immediate return is highest, and an exchange that takes several turns to reach a better-examined answer consumes several times the budget of one that simply answers. Sims (2003) made limited processing capacity an explicit constraint on attention; here a posted access limit creates the constraint, while what the limit costs depends on the user's opportunities. The allowance also binds unevenly. It may be slack for an occasional user and bind for a researcher, developer, or student working through

something difficult, including cases where sustained examination could produce contributions valuable to others. Annex I.D.1a expresses this cost channel in the mode-choice condition.

The unit of metering thus becomes an epistemic choice. A price per token taxes additional deliberation when examination requires additional tokens. A fixed price per resolved task can remove the charge on each extra turn for the user, although the provider still bears the extra computational cost. The renewal architecture below acquires a second job, prior to rewarding shared contributions: an anticipated reward can offset part of the expensive operator's private cost when the user chooses how to engage. An immediate usage subsidy is a further possible instrument, with its own funding requirement. Where validated renewal creates a positive spillover, subsidising its private computational cost has the ordinary externality rationale. The appropriate subsidy still depends on the benefit demonstrated.

Three components follow. A usage charge meters computation; a refundable epistemic deposit contributes cash to a renewal pool; and a validated contribution can earn a delayed royalty for downstream value. Each performs a different job. A cheap query can yield a valuable correction, and an expensive query can leave no reusable contribution. Alongside its financing role, the deposit creates a prospect of a later refund that can increase the expected return to an eligible contribution. The incentive is present at the time of choice even though the money arrives later. A common deposit alone does not favour either operator: the difference comes from their expected validated rewards. An immediate discount would require separate specification and funding under the reserve account in Annex I.D.3.

Design Proposal 1: Epistemic Renewal Architecture. A user chooses whether to enter an arrangement charging for usage and collecting a disclosed epistemic deposit. After the exchange, the user may keep it private or submit an eligible part. Privacy is the default outside the reward path. Submitted material first passes a surprise and relevance screen. A refund or additional reward can vest later, after provenance checks and scrutiny of downstream value by evaluators using different evidence and criteria. Operator choice and sharing remain separate decisions. The epistemic deposit finances the pool; its rate remains a design choice. Annex I.D specifies participation, sharing, corpus renewal, and the accounts; Annex II.D illustrates selected reward and verification rules (Figures 8–11). In that exercise paying for surprise buys the largest flow, but almost half of the paid contributions fail the later value test; a single central verifier pays almost no one; plural delayed validation pays contributions of mean realised value 0.945 spread across many positions.

Delay gives an unusual contribution time to reveal its value. Reuse in independent settings, improved held-out performance, correction of later outputs, or survival across heterogeneous evaluators can provide evidence. Delay also creates selection and governance risks. A pilot must track which contributions are accepted or rejected, who withdraws, and how appeals, privacy losses, and unpaid claims are handled. Delayed scrutiny can obstruct gaming; its effectiveness must be tested.

Ownership determines who captures the residual value. A platform acting as the sole buyer of human variation has monopsony power and can underpay. A cooperative, data union, or commons arrangement gives contributors a continuing interest in the asset they improve (Ostrom 1990). A royalty can reflect value that appears long after submission and continues to accrue. Contributors also need a voice over reuse, validation, and exit.

The creator economy already supports attribution, vesting, and delayed payouts. It also reveals pressure towards popularity and strategic amplification. An epistemic reward system would need to use that infrastructure to track independently demonstrated downstream value, while protecting contributions that begin in low-mass regions of the distribution. Technical feasibility of payment leaves the choice of what deserves payment open.

Verification creates the deepest design risk. A single authority can make the current centre the standard for judging every deviation. Functional plurality requires different tasks, evidence sources, criteria, time horizons, and visible disagreement; a larger panel using one inherited score is not plural. Decisions must be contestable, minority evaluations must remain observable, and verdicts must not propagate automatically across validators. Evaluators must be able to make different mistakes, and the decision rule must keep those differences available for inspection. Annex I.E–F shows why neither pairwise agreement nor a simple consensus score is enough to establish the desired properties.

A credible pilot must examine voluntary participation, privacy costs, validation, strategic behaviour, and the capacity to honour promised payments together. Failure on any of these margins weakens the institutional conjecture even if its accounting identity balances. Annex I.D sets out the detailed reporting and governance requirements.

8. Formal discipline, simulations, and falsification

The formal material separates three kinds of claim. Definitions and exact results cover the measurement coordinates, Page’s identity, variance saturation, corpus mixtures,

and the verifier-joint-law counterexample. Maintained hypotheses concern effort substitution, tail-selective replacement, common-source errors, and population-level Socratic allocation. Design Proposal 1 then specifies a candidate institution for voluntary sharing and delayed rewards. Optimal pricing, welfare effects, and general-equilibrium consequences remain outside the established results.

Annex II retains the computational exercises as implementation records. Annex II.A compares compound corpus regimes (Test 1). Annex II.B combines the analytical equicorrelation result with an unpaired oracle simulation (Test 2). Annex II.C illustrates how an individually challenging intervention can remain population-concentrated. Annex II.D examines selected pricing and validation rules (Test 3) with heterogeneous privacy costs and a simplified sharing decision, without the outside option, strategic validation, provenance, appeal, or intertemporal feasibility. Annex II.E supplies constant-input trajectories of the conditional recurrence; it does not close the institution as a dynamic model.

The depletion account is weakened if AI reliance robustly complements learning on the preregistered tasks; synthetic replacement preserves tails and semantic coverage after marginals and retention are held fixed; provenance similarity fails to predict residual-error dependence. The proposed operator is weakened if its population-level interventions fail to preserve validated alternative paths. The institution is weakened if participation collapses, privacy or validation losses dominate rewards, plural panels fail to improve joint operating characteristics, manipulation cannot be contained, or payments require borrowing against future users. The present dashboard does not adjudicate these conditions. The cost channel of Section 7.2 is weakened on a separate margin: it is tied to a price regime, and it vanishes if the computational gap between examination and assertion closes as inference costs fall.

9. Conclusion: The input economics forgot

Generative AI changes the environment in which new human knowledge is produced. Analysis beginning with separate agents can miss this feedback when it introduces collective knowledge only after individual cognition has been defined. Under the conditions identified here, AI can starve amid textual abundance because its success weakens the independently produced difference on which later models depend.

We claim priority for none of these mechanisms taken separately; learning erosion, recursive degradation, homogenisation, correlated evaluation, and contributor compensation are already present in the emerging literature. The contribution is their connection. A relational

ontology treats knowledge as produced between differently situated people, and the three-coordinate architecture makes quality, variety, and residual-error dependence separately measurable under declared protocols. The two-level Socratic operator then asks whether better conversations preserve different paths across the population, while the prospective institution pays for voluntary sharing under functionally plural validation. Renewal depends on these relations of production and evaluation. The two-bill argument connects them to everyday use: a provider can meter computation accurately while leaving the effect of an exchange on future knowledge outside the price. That gap links the second omission to the third.

This changes the policy question. Compensation for material already used in training addresses a historical claim. Renewal concerns who will continue to learn, record local cases, follow unusual paths, and share discoveries that the model has not supplied. Restitution and renewal both concern enclosure, but the institutions needed for the inherited stock and the future flow differ.

A deposit-funded pool with delayed rewards is one candidate. It could change the return to voluntary sharing while allowing evidence of value to emerge. Verification remains its hardest problem: one authority can turn the current centre into an absorbing standard. Functional plurality requires differences among evaluators and visible disagreement. Participation, privacy, strategic behaviour, administrative cost, and fiscal capacity determine whether the design works. As Annex I.F shows, permanent exclusion requires an additional rule that makes rejection irreversible; a simple consensus score alone does not establish it.

The formal annex establishes limited results: quality alone can omit relevant variation and dependence; common errors can sharply limit variance reduction; and pairwise validator statistics do not determine nonlinear collective decisions. The magnitude of depletion, the optimal price of renewal, and the design's welfare effects remain open. The contribution is an explicit measurement architecture and an institution whose failure conditions can be investigated.

This conditional argument challenges the central presumption of effective accelerationism. e/acc links technological acceleration to greater intelligence and, in some formulations, to greater energy dissipation and entropy production (Beff Jezos and bayes 2022). Here the concern is Shannon uncertainty and a relational input. Faster deployment can raise present cognitive output while depleting the effort, variety, and residual independence required for renewal. Acceleration without institutions that renew human curiosity, creativity, dissent,

and independent judgement can defeat itself. An acceleration that consumes the conditions of future intelligence leads to epistemic liquidation rather than progress.

The wager is simple. Humans remain valuable for a reason that has nothing to do with volume. They continue to encounter the world, learn different things, make partly independent mistakes, and join inherited fragments in ways that no centre can select in advance. AI can help those differences survive examination, and it can help eliminate them before they become visible. Shannon's demon will not starve for lack of words. It will starve if the economy stops renewing the differences that give words new informational value. An economy that prices answers while leaving that renewal outside its accounts may become more capable today and poorer tomorrow.

Annexes

Annex I. Formal discipline: measurements, identities, hypotheses, and a design proposal

The three omissions developed in Sections 2–7 have distinct formal counterparts. Part A specifies what a relational account asks us to measure and what aggregation can establish. Part B states the conditions under which AI use can deplete that measured resource; Part D.1a connects the overlooked computational and epistemic costs to mode choice. Parts C–F then examine the operator, incentives, and validation rules proposed for renewal. Each block begins with the question it addresses; definitions and exact results remain separate from maintained hypotheses and institutional proposals. Annex II provides the associated computational illustrations, with links given where each exercise becomes relevant.

A. Measuring the relational resource and the limits of aggregation

Sections 2–3 argue that a population cannot be understood solely through the performance of its representative member or the precision of one aggregate. To make that claim operational, we first record quality, variety, and error dependence separately. We then examine two results for scalar estimates: the contribution of prediction diversity and the limit imposed by correlated errors. These are measurable implications of the first omission, not a complete mathematical representation of relational knowledge.

A.1. The epistemic state and measurement protocol

The three coordinates introduced in Section 3 answer different questions: does useful performance survive, which cases remain represented, and how far do contributors repeat the same errors? We collect them in a state vector indexed by time. Its components acquire empirical meaning through the protocol below; naming the coordinates does not yet specify their evolution.

$$\mathcal{E}_t = (Q_t, V_t, \rho_t^{\mathrm{err}}) \in [0,1]^3$$

The coordinates $Q, V, \rho^{\mathrm{err}}$ require declared measurements. Task-specific residual-error dependence, ρ^{err}, is distinct from common provenance or source similarity, S^{src}. Their relationship must be assessed. The state becomes meaningful once the task distribution, loss, variety functionals, provenance representation, aggregation rule, and normalisations have been fixed.

A compact display can also be useful. The following constructed product increases with quality and variety and decreases with positive residual-error dependence. We report it as a descriptive summary, with the limits of that interpretation stated in Observation A1.

$$D_t := Q_t V_t (1 - \rho_t^{\mathrm{err}})$$

Observation A1 (limited scalar insufficiency). A statistic increasing in quality alone cannot distinguish two measured states with equal quality and different declared variety or residual-error dependence. It is therefore insufficient for a policy that conditions on either omitted coordinate. This is not a theorem that no scalar sufficient statistic can exist, and $D_t := Q_t V_t (1 - \rho_t^{\mathrm{err}})$ is itself only a constructed summary, not adaptive capacity or welfare.

Measurement protocol. Before observing outcomes, fix a reference-task distribution, bounded loss, corpus or model object, residual unit, variety family, source representation, pair weights, and aggregation rule. One admissible quality normalisation is

$$Q_t := 1 - \frac{\mathbb{E}_{z \sim \Pi_t}[\mathcal{L}(f_t, z)]}{L_{\max}} \in [0,1]$$

$$V_t := A_V\left(H_t^{\mathrm{norm}}, S_t^{\mathrm{supp}}, T_t^{\mathrm{tail}}, C_t^{\mathrm{sem}}\right) \in [0,1]$$

$$\rho_t^{\mathrm{err}} := \left[\sum_{i \neq j} w_{ij,t}\, \mathrm{Corr}_{z \sim \Pi_t}\left(e_{it}(z), e_{jt}(z)\right)\right]_+, \quad S_t^{\mathrm{src}} := \sum_{i \neq j} w_{ij,t}\, s_t^{\mathrm{src}}(i,j)$$

Here the weights are non-negative and sum to one over eligible pairs. The variety aggregator and its component normalisations must be fixed before comparison; entropy, support, tail mass, and semantic coverage remain separately reportable. Zero-variance residual pairs make correlation undefined and must be reported rather than silently imputed. The positive-part dependence diagnostic may be used for the bounded state, while negative residual correlation should be reported separately. The legacy dashboard complexity statistic remains a secondary constructed sensitivity diagnostic and is not part of the measurement protocol. The reference-task distribution must remain fixed across a comparison, or changes in it must be reported separately; otherwise a change in measured quality can reflect a change in the tasks being asked.

A.2. What aggregation delivers and when it saturates

Scalar aggregation still has a precise role. Section 3 uses Page's identity to distinguish individual prediction error from diversity around the crowd mean. Section 4 then shows why more contributors need not deliver proportionately greater precision when their errors co-vary. The following results formalise those two claims under their respective assumptions. Annex II.B places them alongside the shared-oracle illustration; Figure 4 plots the analytical saturation result directly.

For predictions x_i of a scalar quantity θ with crowd mean $\bar{x}$, squared error decomposes exactly as:

$$(\bar{x} - \theta)^2 = \frac{1}{N}\sum_i (x_i - \theta)^2 - \frac{1}{N}\sum_i (x_i - \bar{x})^2.$$

Crowd error equals average individual error minus prediction diversity (Page 2007). The identity applies to a common scalar target under squared loss. Contested domains may lack a unique target of this kind. The identity alone does not imply that lower effort increases individual error or that common-source reliance lowers diversity; those causal statements require an effort-to-error relation and a specified joint distribution.

Let $N \geq 2$ signals have common variance σ^2 and equicorrelation $-\frac{1}{N-1} \leq \rho \leq 1$. Then:

$$\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{N}[1 + (N-1)\rho],$$

$$N_{\text{eff}} = \frac{N}{1 + (N-1)\rho}, \qquad \text{provided } 1 + (N-1)\rho > 0.$$

Proposition A2 (variance saturation). Under equal weights, common variance, equicorrelated residuals, and fixed $\rho > 0$, the variance-equivalent sample size N_{eff} approaches $1/\rho$ as N grows. This is a ceiling on variance reduction for that equal-weight mean, not a general ceiling on information, semantic coverage, or effective independence. When every residual error is zero, residual correlation and this correction are unnecessary.

B. Depletion dynamics and conditional hypotheses

The second omission concerns the costs of AI use, including changes to the resource on which later use depends. Section 4 identifies effort substitution, tail-selective replacement, and common-error provenance; Section 5 explains why their effects cannot be judged from one

coordinate alone. The recurrence below records possible movements of the state just defined, while H1–H3 state the conditional signs to be investigated. Annex II.A–B illustrates corpus composition and common-source reliance; Annex II.E applies the recurrence with supplied inputs. These exercises do not estimate the feedback functions. Part D.1a subsequently formalises how computational costs and rationed access can influence the choice between conversational modes; it does not equate energy use with epistemic value.

Potential users first compare participation with their outside option and, conditional on participating, choose reliance and an operator mode. Exchanges then occur; sharing is chosen after the contribution is observed; earlier contributions complete validation and vest; and the corpus and account update. Let $u_t = (a_t, h_t, \xi_t, F_t, \Sigma_t, S_t^{\mathrm{src}})$. Here a records AI reliance, h fresh human inflow, ξ synthetic replacement, F validated structured novelty, and Σ population-level Socratic dispersion; the final input records source similarity separately from error dependence. The learning response ϑ(a) connects reliance to retained learning. The μ coefficients represent attrition; the γ and λ coefficients scale the indicated additions and losses; ε terms allow shocks. After fixing their domains, the response, supplied inputs, and shock law, the dashboard-style recurrence can be written as

$$\begin{aligned} Q_{t+1} &= \Pi_{[0,1]}\left((1-\mu_Q)Q_t + \gamma_Q \vartheta(a_t) + \varepsilon_{t+1}^{Q}\right), \\ V_{t+1} &= \Pi_{[0,1]}\left((1-\mu_V)V_t + \gamma_h h_t + \gamma_F F_t - \lambda_V \xi_t + \varepsilon_{t+1}^{V}\right), \\ \rho_{t+1}^{\mathrm{err}} &= \Pi_{[0,1]}\left((1-\mu_\rho)\rho_t^{\mathrm{err}} + \gamma_\rho S_t^{\mathrm{src}} - \lambda_\rho \Sigma_t + \varepsilon_{t+1}^{\rho}\right). \end{aligned}$$

Projection keeps reported coordinates within the unit cube; it does not identify the controls, coefficients, causal directions, or policy response. The companion dashboard supplies constant inputs and zero shocks. Its trajectories are conditional calculations, not an endogenous Markov closure of the institution.

Maintained hypotheses for preregistered tests. The paper's three dynamic claims are

$$\text{H1: } \frac{\partial Q_{t+1}}{\partial a_t} < 0, \qquad \text{H2: } \frac{\partial V_{t+1}}{\partial \xi_t} < 0, \qquad \text{H3: } \frac{\partial \rho_{t+1}^{\mathrm{err}}}{\partial S_t^{\mathrm{src}}} > 0$$

H1 is conditional on AI substituting for effort rather than complementing learning. H2 is conditional on replacement removing tail or semantic coverage rather than adding independently verified material. H3 is conditional on common provenance producing common task errors; model diversity, independent verification, or negative error correlation can reverse or eliminate it. These are falsifiable sign restrictions, not consequences of the coordinate definitions. The strict derivatives refer to differentiable interior regions with other inputs held

fixed. At a projected boundary a response can be zero. In the displayed recurrence, H1 also requires the learning response to fall with reliance; the signs are imposed hypotheses rather than findings generated by the recurrence.

Fixing every response function and shock law would define a mathematical process, but the paper neither estimates those objects nor treats that closure-by-assumption as a contribution. The exact results elsewhere in this annex are conditional on declared measures and maintained assumptions; causal and policy comparative statics require empirical restrictions beyond the annex.

If the recurrence is later fitted, deterministic fixed points, projected boundary states, and stochastic invariant measures can be analysed under the usual regularity conditions. No stationary state is interpreted here as an equilibrium or welfare optimum.

C. Mirror and Socratic operators at two levels

Section 6 asks whether a challenge improves one conversation while sending many users towards the same alternative. That distinction requires us to separate what can be assessed when a candidate is offered from what becomes known only after scrutiny. We therefore define surprise and initial relevance first, then downstream value and marginal independence. Annex II.C illustrates the individual–population distinction in two calibrations, without treating its geometric dispersion statistic as a measure of independent errors.

The two densities below compare the current model with a declared reference. The reference determines what counts as underrepresented; surprise is measured relative to that comparison, rather than as an absolute measure of originality.

Let $\zeta \in \mathcal{C}$ denote observable context. Let $p_t(x \mid \zeta)$ and $p_{\mathrm{ref}}(x \mid \zeta)$ be densities with respect to one declared dominating measure, with $p_t(x \mid \zeta) > 0$ on the eligible support. Write $[u]_+ := \max\{u, 0\}$. For a fixed normalisation scale $L_z > 0$ define:

$$\bar{I}_t(x \mid \zeta) = \min\left\{1, \frac{[\log\{p_{\mathrm{ref}}(x \mid \zeta)/p_t(x \mid \zeta)\}]_+}{L_z}\right\} \in [0,1]$$

Let $r_t^{\mathrm{pre}}(x, \zeta) \in [0,1]$ denote relevance observable at selection, $w_t^{\mathrm{post}}(x) \in [0,1]$ downstream marginal value estimated after external scrutiny, and $\Delta_t^{\mathrm{ind}}(x) \in [0,1]$ attributable marginal independence after provenance and de-duplication. Realised structured novelty is:

$$z_t(x \mid \zeta) = \bar{I}_t(x \mid \zeta)\, r_t^{\mathrm{pre}}(x, \zeta)\, w_t^{\mathrm{post}}(x)\, \Delta_t^{\mathrm{ind}}(x) \in [0,1]$$

Ex-ante selection and ex-post validation must be separated. A Socratic candidate may be selected using $\bar{I}_t$ and r_t^{pre}, which are available at the exchange. It becomes a validated Socratic contribution only after w_t^{post} and Δ_t^{ind} are estimated. The operator therefore cannot select directly on future downstream value. The dashboard follows this timing only to the extent that hidden downstream value is used after selection; provenance and marginal-independence validation remain implementation requirements.

A Mirror operator favours high-probability continuations under p_t. Repeated use can narrow relevant support and raise cross-user similarity. An individually Socratic operator surfaces an underrepresented candidate that passes the ex-ante surprise and relevance screens. A population-level Socratic operator adds a further condition: the candidates injected across users retain positive dispersion and imperfect correlation. One model can give a million users the same strong counterargument. Each conversation may then be individually Socratic while the population receives one voice.

The dashboard's cross-user geometric dispersion is an implementation statistic. It is neither residual-error dependence nor source similarity. In the exact mixture identity below, validated contributions first determine a population law; declared residual-error and source-similarity kernels are then evaluated separately, including all within- and cross-component terms.

D. A voluntary architecture for delayed renewal rewards

Section 7 turns the third omission into an institutional question: how can a provider pay for contributions whose value emerges later, while users retain the option to keep an exchange private? Design Proposal 1 is formalised in three steps: individual participation and sharing, the resulting corpus inflow, and the account that finances promised payments. Annex II.D illustrates selected incentive and validation rules; the fuller participation and reserve system below remains a proposal for implementation and testing.

D.1. Individual timing and the joint validation and payment law

The distinction between disclosure and payment in Sections 7.1–7.2 makes timing consequential. A user chooses whether to participate before seeing the contribution, decides whether to share afterwards, and may receive payment only after validation. The joint law below records acceptance, delay, and payment together, so that a high chance of acceptance is not mistaken for a high present value of the reward.

Fix a maximum resolution horizon $H \in \mathbb{N}_+$. The pre-exchange state and policy are observable. User i first decides whether to participate, and conditional on participation chooses $m_{i\tau} \in \{M, S\}$ before (x, ι) is drawn from $G_{i\tau}(dx, d\iota \mid m_{i\tau}, \mathcal{S}_\tau, \psi_\tau)$. After observing the pair, the user chooses whether to share. Sharing can impose both a direct cost and a privacy loss. The symbols M and S denote Mirror and Socratic modes. The draw contains the candidate contribution x and its associated metadata ι. The pre-exchange state S and policy ψ collect the information used in the declared contribution and validation laws; they do not supply an estimated law of motion.

Conditional on a submitted pair, architecture g is a stochastic kernel

$$\mathcal{K}^g_{i\tau}(dy, d\ell, dj, d\varpi \mid x, \iota, \mathcal{S}_\tau, \psi_\tau)$$

$$(y, j) \in \{0,1\}^2, \quad \ell \in \{1, \dots, H\}, \quad \varpi \in [0, \overline{\varpi}]$$

Here y is validation, ℓ is the resolution and vesting date, j is non-default conditional on an accepted claim, and ϖ is the additional prize. The prize marginal is degenerate at the publicly committed schedule $\varpi = \varpi_{i\tau}(x, \iota, \mathcal{S}_\tau, \psi_\tau) \in [0, \overline{\varpi}]$; consequently the platform cannot select a prize after observing validation. Set $j = 1$ when $y = 0$ by convention. The kernel otherwise permits dependence among validation, timing, and non-default. In the payment expressions, capital Y and J are the realised validation and non-default indicators, and T is the realised delay l. The post-exchange metadata used in the account are the realised counterpart of ι.

The acceptance probability and the probability of a positive payment at horizon ℓ are different marginals of this same kernel:

$$\begin{aligned} q^g_{i\tau}(x, \iota) &:= \int y \, \mathcal{K}^g_{i\tau}(dy, d\ell, dj, d\varpi \mid x, \iota, \mathcal{S}_\tau, \psi_\tau), \\ p^{\text{pay},g}_{i\tau,\ell}(x, \iota) &:= \int yj \, 1\{\varrho d_\tau + \varpi > 0\} \, \mathcal{K}^g_{i\tau}(dy, \{\ell\}, dj, d\varpi \mid x, \iota, \mathcal{S}_\tau, \psi_\tau). \end{aligned}$$

Assume $d_\tau \in [0, \infty)$, $\varrho \in [0,1]$, $p_e \in [0, \infty)$, $e_i(m) \in [0, \infty)$, $b_m \in \mathbb{R}$, $\overline{\varpi} < \infty$. With deposit-refund fraction ϱ, the current charge and realised payment are

$$\begin{aligned} C_{i\tau} &= p_e e_i(m_{i\tau}) + d_\tau, \\ R_{i\tau,t} &= \chi_{i\tau}(x, \iota) \, 1\{T_{i\tau} = t - \tau\} \, Y^g_{i\tau} J_{i\tau} \, (\varrho d_\tau + \varpi_{i\tau}), \qquad t > \tau. \end{aligned}$$

In the payment, the indicator χ records whether the user shares; its optimal value χ* is derived below. Assume $\delta_i \in [0,1)$. The finite-horizon conditional present value is the kernel integral

$$PV_{i\tau}^{g}(x,\iota) = \sum_{\ell=1}^{H} \delta_i^{\ell} \int yj(\varrho d_\tau + \varpi)\, \mathcal{K}_{i\tau}^{g}(dy, \{\ell\}, dj, d\varpi \mid x, \iota, \mathcal{S}_\tau, \psi_\tau) \in [0, \varrho d_\tau + \overline{\varpi}]$$

Let direct sharing cost and subjective privacy loss be denominated in the same utility units as the expected payment, with $\mathrm{k}_i^{\mathrm{sh}}(x,\iota), \kappa_i^{\mathrm{priv}}(x,\iota) \in [0,\infty)$. Under risk neutrality, the sharing rule and its mode-specific option value are

$$\chi_{i\tau}^{*}(x,\iota;g) = 1\{PV_{i\tau}^{g}(x,\iota) \geq k_i^{\mathrm{sh}}(x,\iota) + \kappa_i^{\mathrm{priv}}(x,\iota)\}$$

$$\Phi_{i\tau}(m,g) = \int \left[PV_{i\tau}^{g}(x,\iota) - k_i^{\mathrm{sh}}(x,\iota) - \kappa_i^{\mathrm{priv}}(x,\iota)\right]_{+} G_{i\tau}(dx, d\iota \mid m, \mathcal{S}_\tau, \psi_\tau)$$

Let $\kappa_i^{S} \in [0,\infty)$ be additional Socratic friction and let the non-participation payoff satisfy $U_{i\tau}^{\mathrm{out}} \in \mathbb{R}$. Then

$$\begin{aligned} U_{i\tau}(m;g) &= b_m - \kappa_i^{S} 1\{m = S\} - p_e e_i(m) - d_\tau + \Phi_{i\tau}(m,g), \\ \overline{U}_{i\tau}(g) &:= \max_{m\in\{M,S\}} U_{i\tau}(m;g), \\ \pi_{i\tau}^{*}(g) &= 1\{\overline{U}_{i\tau}(g) \geq U_{i\tau}^{\mathrm{out}}\}, \qquad m_{i\tau}^{*} \in \operatorname{argmax}_m U_{i\tau}(m;g). \end{aligned}$$

D.1a. The compute wedge and rationed access

The cost argument in Section 7.2 concerns a barrier that can arise before any contribution is shared. When examination requires more inference than a direct answer, metering or a binding allowance can change the user's choice of operator. The comparison below inserts that computational gap into the utility already specified and distinguishes the provider's cost from the user's scarce allowance.

The mode-specific inference requirement $e_i(m)$ is already a primitive of the charge $C_{i\tau}$. The comparison rests on the cost gap

$$\Delta e_i := e_i(S) - e_i(M) > 0.$$

Examination is assumed to consume more computation than assertion in this comparison; the definition of a Socratic intervention alone does not impose that ranking. Substitution into $U_{i\tau}(m;g)$ gives the condition for choosing the Socratic mode:

$$(b_S - b_M) + (\Phi_{i\tau}(S,g) - \Phi_{i\tau}(M,g)) \geq \kappa_i^{S} + p_e \Delta e_i.$$

The metered price enters on the same side as the Socratic friction κ_i^{S}. The consequences appear on both sides of the exchange, through the provider's cost and the user's scarce allowance.

On the supply side, let the platform serve at a flat subscription p_{sub} with expected computational cost $c_e \mathrm{E}[e_i(m_{i\tau}^*)]$. Its unit cost c_e is distinct from the user price p_e. Holding subscription revenue, task mix, and other costs fixed, its per-user margin decreases as the Socratic share rises. A platform choosing defaults, interface affordances, and response policies to maximise that margin therefore selects weakly towards the cheaper operator. The pressure comes from computational cost, separately from the attention channel.

On the demand side, replace the metered charge by an allowance. The user faces

$$\sum_{\tau} e_i(m_{i\tau}) \leq B_i, \quad \lambda_i \geq 0,$$

where λ_i is the constraint's multiplier and is zero when the allowance is slack. The condition for choosing the Socratic mode becomes

$$(b_S - b_M) + (\Phi_{i\tau}(S, g) - \Phi_{i\tau}(M, g)) \geq \kappa_i^S + \lambda_i \Delta e_i.$$

The allowance and the user's available tasks and payoffs determine λ_i. The comparison with a single multiplier describes the relaxation in which the allowance can be divided across tasks, or a solution supported by one common shadow price; with indivisible choices between M and S, one multiplier need not reproduce the optimal set of Socratic tasks. Holding the benefit side fixed, Socratic choice becomes weakly less attractive as this shadow price rises. More intensive use may make a cap bind, but intensity alone does not rank shadow prices across heterogeneous users. Metering and rationing differ in who bears the cost gap and how it is distributed. Both comparisons follow from the declared utility and the maintained gap, with the other payoff terms held fixed; they establish a conditional direction rather than a magnitude. Posted allowances and charging rules can be observed directly. The task-specific gap Δe_i requires measured or explicitly assumed inference requirements and need not be disclosed by a provider.

The acceptance probability alone does not determine expected payment: resolution dates, non-default, and prize amounts also matter. The displayed inequalities show only how an announced reward can change participation and sharing under the maintained preferences. They do not establish truthful reporting, welfare improvement, privacy adequacy, or a feasible reward level. Equality selects participation and sharing as stated tie-breaking conventions.

D.2. From individual contributions to the corpus

An incentive matters for renewal only if it changes what enters the shared resource. This block connects the decisions in D.1 to the corpus composition discussed in Section 4 and the renewal proposal in Section 7. We construct the distribution of validated submissions, combine it with retained material and other inflows, and evaluate the resulting corpus. Source similarity and residual-error dependence remain separate measurements throughout this construction. The causal response functions remain open in Part B. With delayed validation, a cohort's eventual contribution law and the flow available for incorporation today need separate dates. The expressions below define the former and an abstract mixture update; routing completed cohorts into that update remains to be specified in a dynamic implementation.

Let the period-t potential-user space carry probability measure $\nu_t(\mathcal{I}_t) = 1$. The contribution kernel describes potential draws conditional on the best platform mode; the participation indicator below removes non-participants. Conditional on the declared behavioural and validation primitives, the potential cohort law is

$$\Gamma_t^g(di, dx, d\iota, dy, d\ell, dj, d\varpi) = \nu_t(di)\, G_{it}(dx, d\iota \mid m_{it}^*, \mathcal{S}_t, \psi_t)\, \mathcal{K}_{it}^g(dy, d\ell, dj, d\varpi \mid x, \iota, \mathcal{S}_t, \psi_t)$$

The validated submission rate, the probability law of validated contributions, and the normalised structured-novelty flow are

$$\begin{aligned} \eta_t^g &:= \int \pi_{it}^*(g)\chi_{it}^*(x,\iota;g)y\, d\Gamma_t^g \in [0,1], \\ H_t^g(A) &:= \frac{1}{\eta_t^g}\int \pi_{it}^*(g)\chi_{it}^*(x,\iota;g)y1\{x \in A\}\, d\Gamma_t^g, \qquad \eta_t^g > 0, \\ F_t^g &:= \int \pi_{it}^*(g)\chi_{it}^*(x,\iota;g)y\, z_t(x \mid \zeta_{it})\, d\Gamma_t^g \in [0,1]. \end{aligned}$$

If $\eta_t^g = 0 \Rightarrow \omega_{gt} = 0$ and the otherwise undefined validated-contribution distribution receives zero corpus weight. Let P_t be the retained-corpus law, let H_t^0 be ordinary fresh human inflow, let H_t^g be the validated law above, and let A_t^{syn} be synthetic inflow. Write $P_t^0 = P_t$, $P_t^h = H_t^0$, $P_t^g = H_t^g$, $P_t^s = A_t^{\text{syn}}$ and choose measurable state-policy weights $\omega_t = (\omega_{0t}, \omega_{ht}, \omega_{gt}, \omega_{st}) \in \Delta_4$. The corpus transition is the exact mixture

$$P_{t+1} = \sum_{a \in \{0,h,g,s\}} \omega_{at}\, P_t^a$$

This block chooses a particular pairwise-kernel protocol. The bilinear variety identity below applies to that choice; it does not hold automatically for entropy, support, or every admissible variety measure listed in A.1.

Augment contributions with provenance and task-evaluation labels. Fix, rather than infer implicitly, a reference quality score, a symmetric normalised variety kernel, a symmetric residual-error dependence kernel, and a symmetric source-similarity kernel with domains

$$q^*: \mathcal{X} \to [0,1], \quad d_V, k_{\mathrm{err}}, k_{\mathrm{src}}: \mathcal{X}^2 \to [0,1]$$

$$\begin{aligned} Q(P) &:= \int q^*(x)\, P(dx), \\ V(P,R) &:= \iint d_V(x,x')\, P(dx) R(dx'), \\ \rho^{\mathrm{err}}(P,R) &:= \iint k_{\mathrm{err}}(x,x')\, P(dx) R(dx'), \\ S^{\mathrm{src}}(P,R) &:= \iint k_{\mathrm{src}}(x,x')\, P(dx) R(dx'). \end{aligned}$$

$$Q_t = Q(P_t), \qquad V_t = V(P_t, P_t), \qquad \rho_t^{\mathrm{err}} = \rho^{\mathrm{err}}(P_t, P_t), \qquad S_t^{\mathrm{src}} = S^{\mathrm{src}}(P_t, P_t)$$

Bilinearity gives the exact component expansion

$$\begin{aligned} Q_{t+1} &= \sum_a \omega_{at}\, Q(P_t^a), \\ V_{t+1} &= \sum_{a,b} \omega_{at}\, \omega_{bt} V\left(P_t^a, P_t^b\right), \\ \rho_{t+1}^{\mathrm{err}} &= \sum_{a,b} \omega_{at}\, \omega_{bt} \rho^{\mathrm{err}}\left(P_t^a, P_t^b\right), \\ S_{t+1}^{\mathrm{src}} &= \sum_{a,b} \omega_{at}\, \omega_{bt} S^{\mathrm{src}}\left(P_t^a, P_t^b\right). \end{aligned}$$

The mixture identities map individual sharing and validation into a corpus distribution and then into declared measurements. They leave causal responses unspecified, and source similarity and residual-error dependence remain different functionals: high common provenance need not generate common errors on every task, and common errors can arise without shared provenance. The dashboard dispersion statistic remains a descriptive proxy and cannot be substituted for either functional without validation.

D.3. The budget and a conservative reserve envelope

Delayed rewards create commitments that a current-period balance can conceal. Section 7 requires the institution to honour those commitments. The account below therefore follows both cash and unresolved claims, reserving their full face value before allowing withdrawals.

This supplies an accounting check on the proposal, separate from whether users join or whether the institution improves welfare.

Let $B_t \geq 0, \quad \mathcal{M}_t \in [0,\infty), \quad r_{B,t} > -1, \quad A_t, C_t^{\mathrm{op}}, W_t \in [0,\infty)$, where $\mathcal{M}_t$ is realised participant mass; in the order listed, the other symbols are the cash balance, its rate of return, external contributions, operating costs and withdrawals. Define potential-user mass and realised participant mass by

$$\overline{\mathcal{M}}_t \in [0,\infty), \qquad \mathcal{M}_t := \overline{\mathcal{M}}_t \int_{\mathcal{I}_t} \pi_{it}^*(g)\, \nu_t(di)$$

Aggregate realised payout is

$$P_t^{\mathrm{out}} := \sum_{\tau<t} \overline{\mathcal{M}}_\tau \int_{\mathcal{I}_\tau} \pi_{i\tau}^*(g) \chi_{i\tau}^*\big(X_{i\tau}, I_{i\tau}^{\mathrm{post}}; g\big) 1\{\tau + T_{i\tau} = t\} Y_{i\tau}^g J_{i\tau} (\varrho d_\tau + \varpi_{i\tau})\, \nu_\tau(di)$$

A conservative accounting policy reserves the full face value of every unresolved submitted claim, including claims later rejected:

$$L_t := \sum_{\tau<t} \overline{\mathcal{M}}_\tau \int_{\mathcal{I}_\tau} \pi_{i\tau}^*(g) \chi_{i\tau}^*\big(X_{i\tau}, I_{i\tau}^{\mathrm{post}}; g\big) 1\{\tau + T_{i\tau} \geq t\} (\varrho d_\tau + \varpi_{i\tau})\, \nu_\tau(di)$$

The cash-account transition is

$$B_{t+1} = \big(1 + r_{B,t}\big) B_t + d_t \mathcal{M}_t + A_t - C_t^{\mathrm{op}} - P_t^{\mathrm{out}} - W_t$$

The mechanism may issue new reward commitments only when the following reserve inequality has a non-negative right-hand side, and it may withdraw no more than the resulting surplus:

$$0 \leq W_t \leq \big(1 + r_{B,t}\big) B_t + d_t \mathcal{M}_t + A_t - C_t^{\mathrm{op}} - P_t^{\mathrm{out}} - L_{t+1}$$

Accounting Lemma D1 (sufficient reserve invariant). If $B_0 \geq L_0$, the reserve rule is enforced with a non-negative right-hand side in every period, and accepted claims are honoured, then $B_t \geq L_t \geq 0$ for every t pathwise. Substitution gives $B_{t+1} \geq L_{t+1}$ and induction proves the invariant. If operating costs or a negative return make that right-hand side negative, the rule cannot be met and the invariant can fail even without new commitments. This result shows only that the chosen conservative rule can prevent the account from promising more face value than it reserves. It does not prove user participation, efficient pricing, administrative viability, robustness to manipulation, or economic feasibility.

A welfare change is not cash. No welfare-linked transfer enters the account unless a separately justified money conversion is supplied. The paper estimates neither that conversion nor a welfare function, and therefore makes no welfare-improvement claim.

Formal status. The exact objects in this annex are definitions, accounting identities, or conditional mixture expansions given the measurable primitives $G, \mathcal{K}, q^*, d_V, k_{\text{err}}, k_{\text{src}}$. The sign restrictions are maintained hypotheses, and Design Proposal 1 is an institutional conjecture. No claim is made that the manuscript specifies or estimates a complete Markov economy, an equilibrium participation system, an optimal mechanism, or a Pigouvian levy.

Pilot reporting and governance requirements. Before validation, commit the reward schedule and preserve a meaningful outside option. Minimise and disclose privacy costs; distinguish provenance similarity from realised error dependence; measure validator operating characteristics and joint decisions; reserve against promised payments; and permit appeal and contestation. A pilot must report false acceptance and rejection, attrition, privacy losses, appeal outcomes, provenance failures, payment concentration, and unresolved liabilities. It must also examine duplication, Sybil identities, collusion, and strategic poisoning. These requirements supplement the accounting identities: a balanced account alone cannot establish voluntary participation or successful validation.

E. Plural verification and dependence among verifiers

The verification problem at the end of Section 7 cannot be settled by counting evaluators or measuring their pairwise correlation. Payment depends on their joint decision under a quorum rule. The counterexample below shows why two panels with identical marginal voting probabilities and pairwise correlations can accept a contribution with different probabilities. It identifies what a formal account of functional plurality must specify beyond the variance calculation in A.2.

Let $A_j(x, \theta) \in \{0,1\}, \ j = 1, \dots, n_g$ be verifier votes for contribution x under task/value state θ. Architecture g must specify the full conditional joint law $\mathcal{J}_g\left(A_1, \dots, A_{n_g} \mid x, \theta\right)$. For quorum k, acceptance is

$$q^g(x, \theta) = \Pr_g\left(\sum_{j=1}^{n_g} A_j \geq k \,\middle|\, x, \theta\right)$$

Proposition E1 (pairwise correlation is insufficient). The expression $\frac{n_g}{1+(n_g-1)\rho_g}$ is a variance-equivalent sample size for a linear equal-weight average under equicorrelation;

it does not characterise a nonlinear quorum or verifier accuracy. For $n_g = 3$ and $k = 2$ the uniform laws on $\{000,011,101,110\}$ and $\{111,100,010,001\}$ have identical Bernoulli-$1/2$ marginals and zero pairwise correlations, but acceptance probabilities $3/4$ and $1/4$. Functional plurality therefore requires the joint vote law and operating characteristics, not a copied effective-sample-size formula.

F. A naive consensus rule and its limits

Sections 2 and 7 warn that early agreement can suppress an unusual contribution before its value becomes visible. A simple score that rises with support and falls with conflict may appear to capture that danger. The recurrence below lets us inspect the claim directly: temporary suppression follows in some cases, but irreversible exclusion does not follow from this rule alone. The negative result marks the additional mechanism a model of permanent loss would need.

Let candidate relation (i, j) receive weight w_t^{ij} updated by:

$$w_{t+1}^{ij} = w_t^{ij} + \alpha\, \mathrm{support}_t^{ij} - \beta\, \mathrm{conflict}_t^{ij}.$$

A high-value corner relation can begin with low support and high conflict with the current smooth reconstruction, so the rule can suppress it before downstream validation reveals its value. The displayed recurrence is not absorbing: for $\alpha = \beta = 1$ and $w_0 = 1$, $(\mathrm{support}_0, \mathrm{conflict}_0) = (0,1)$ gives $w_1 = 0$, while $(\mathrm{support}_1, \mathrm{conflict}_1) = (1,0)$ gives $w_2 = 1$. Recovery is immediate, and the equation contains no verifier-dependence parameter. Absorption requires a separately imposed irreversible active/deletion state; it is not formally equivalent to perfectly correlated verifiers.

Annex II. Simulation: computational mirrors and boundary illustrations

The exercises follow the questions raised in the main text: what happens to a repeatedly refitted corpus, what shared reliance does to aggregation, whether challenges remain plural across users, and what different reward rules purchase. A final recurrence shows how the measured coordinates can move in different directions. Each exercise points back to the relevant argument and formal block. Readers can use these links to move between the narrative, assumptions, and numerical output without treating an illustration as an empirical test. Every table and figure in this annex is produced by the interactive application (https://microeconomics.shinyapps.io/AI_Starves/) with seed 1: on each page, the tab in the article reruns the reported calibration with the button Run with the article's values. Where a pattern depends on the realisation, the text reports how it behaves across seeds. Computations run in R with base graphics; the application uses shiny and flexdashboard, and its file lists every package.

Scope of the simulations. The exercises below illustrate the definitions, hypotheses, and institutional proposal in Annex I under selected parameterisations. They show how the specified objects can be computed and how they behave within the coded environments. They do not estimate the preregistered measurement system, participation and privacy selection, the full validation law, or the accounting envelope. Causal mechanisms, equilibrium, and welfare comparisons therefore remain outside their demonstrated scope. The paper's central claims draw on the exact results in Annex I and the evidence cited in Section 4; these simulations supply illustrations of assumptions and counterexamples. The following sections describe what each exercise implements and the limits specific to its design.

A. Test 1: Feeding the corpus

Section 4 describes the risk that generated material displaces the observations anchoring rare cases. This exercise makes that concern visible through repeated fitting and sampling in a finite token vocabulary. It is linked to the variety coordinate in Annex I.A.1 and the replacement hypothesis in Annex I.B. The four regimes below compare alternative ways of feeding and retaining the corpus.

A generator is fitted, sampled, and refitted under four regimes. Pure replacement lets synthetic output displace the prior corpus. Independent accumulation adds synthetic output and genuinely fresh human observations. Fixed-window accumulation uses the same

independent mix but resamples to a constant corpus size; this changes retention and resampling as well as mass, so it does not isolate corpus size. The coded 'correlated' stream also changes marginal composition while changing repetition, so it does not isolate dependence. Tracked quantities are normalised token entropy, rare-tail mass, vocabulary support, and held-out cross-entropy against the original human distribution. The exercise therefore compares compound regimes. The interpretation below returns to the distinctions needed for causal attribution.

Benchmark calibration and results. The first reported run provides a benchmark under one selected parameterisation. It is illustrative; other parameterisations remain possible. The vocabulary contains 200 token types and the corpus evolves for 25 generations. Each generation adds 1,000 human and 3,000 synthetic tokens. The compression parameter is 1.04, shared reliance in the correlated inflow is 0.80, and the reported entries are final-generation means across ten Monte Carlo histories (seed 1; Table 1). The human distribution is Zipf with exponent 1.10 over the 200 types, and the first corpus holds 10,000 tokens drawn from it. Each generation fits the observed frequencies and sharpens them by raising them to the power of the compression parameter and renormalising. Rare-tail mass is the share of the rarest fifth of types; held-out cross-entropy is measured against the human distribution. Fixed-window accumulation resamples the pooled corpus back to 10,000 tokens. In the coded 'correlated' stream every nominally human token is drawn from the mixture $0.20 \times \text{human} + 0.80 \times \text{model}$.

Table 1. Final-generation variety diagnostics across the four corpus regimes

Regime	Normalised entropy	Rare-tail mass	Vocabulary support	Held-out loss
Correlated accumulation	0.6465	0.0179	1.000	3.9020
Fixed-window accumulation	0.6258	0.0185	0.920	4.2359
Independent accumulation	0.6669	0.0205	1.000	3.8811
Replacement	0.1322	0.0003	0.084	13.2420

Note: Final-generation means across ten Monte Carlo histories, seed 1. Parameters: vocabulary=200; generations=25; compression=1.04; shared reliance=0.80; human tokens per cycle=1,000; synthetic tokens per cycle=3,000; variant=mixture.

Replacement produces a clear collapse across all four diagnostics. Normalised entropy falls from approximately 0.70 to 0.1322, vocabulary support contracts to 8.4 per cent, and rare-tail mass is almost eliminated. Held-out cross-entropy rises to 13.242, more than three times the level observed under the accumulation regimes. The trajectories in Figure 1 and the rank-

frequency profile in Figure 2 show that the centre of the distribution survives longer than the tail. Both axes of Figure 2 are logarithmic, and values at the plotting floor are token types absent or effectively absent in the realised corpus. A small group of recurrent forms remains visible while most low-probability material disappears.

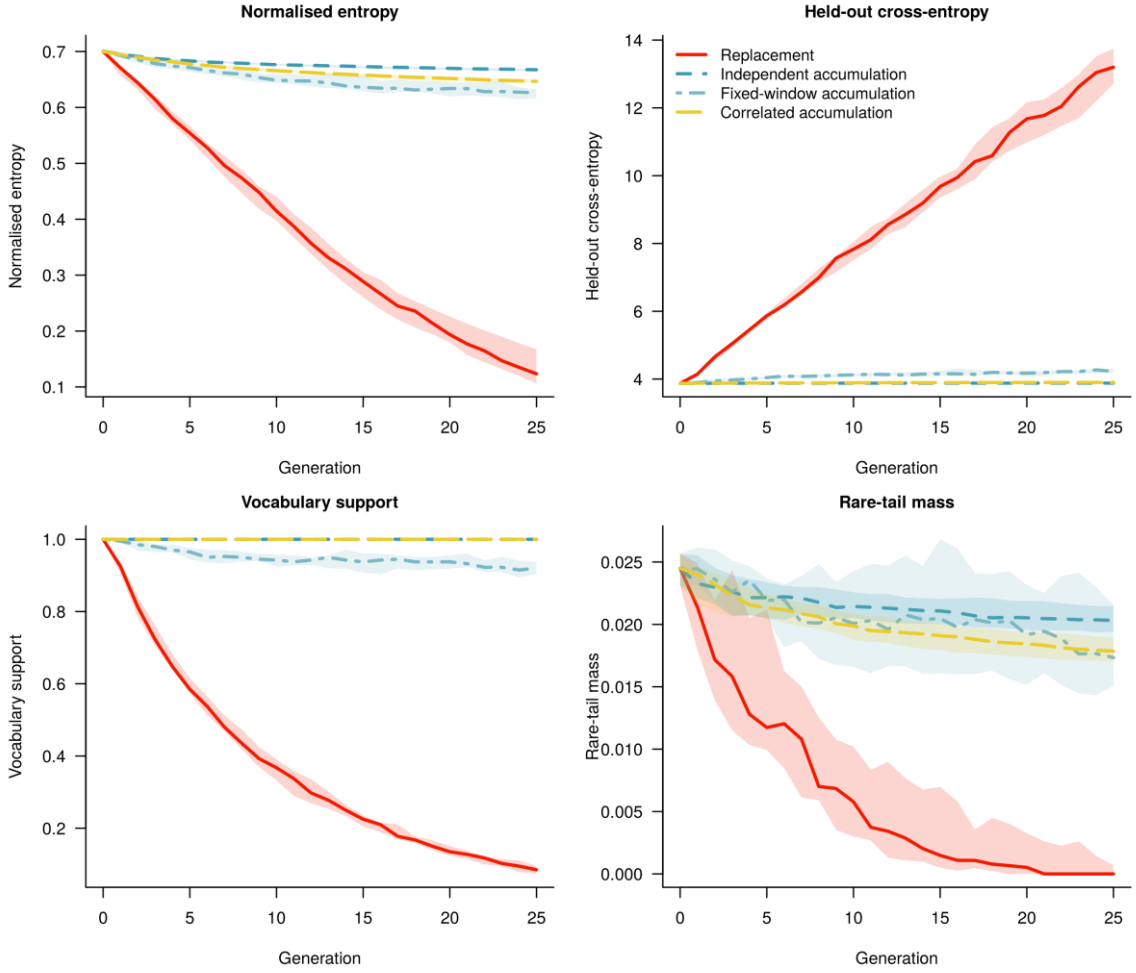
Independent accumulation has the strongest reported diagnostics in this calibration. It ends with normalised entropy of 0.6669, full vocabulary support, rare-tail mass of 0.0205, and held-out loss of 3.8811. The coded 'correlated' and independent streams also differ in marginal composition, not only dependence. Their numerical gap therefore cannot be attributed causally to correlation or independence.

The fixed-size comparison changes corpus mass, retention, and resampling together; it therefore does not isolate the mechanical effect of a growing corpus. In the reported run the diagnostics cross: correlated accumulation preserves full support and lower held-out loss, while fixed-size resampling retains slightly more rare-tail mass, 0.0185 compared with 0.0179. Across 30 seeds this crossing appears in 14; in the others correlated accumulation ranks above fixed-size resampling on all four diagnostics. Whether the four diagnostics order these two regimes alike therefore depends on the realised world.

The differences belong to four compound coded regimes. Attributing them to replacement, accumulation, or independence would require a paired factorial design that holds corpus mass and one-token marginals fixed while manipulating retention, replacement, and dependence separately. The ten histories document the comparison, while providing only a limited basis for quantifying Monte Carlo uncertainty.

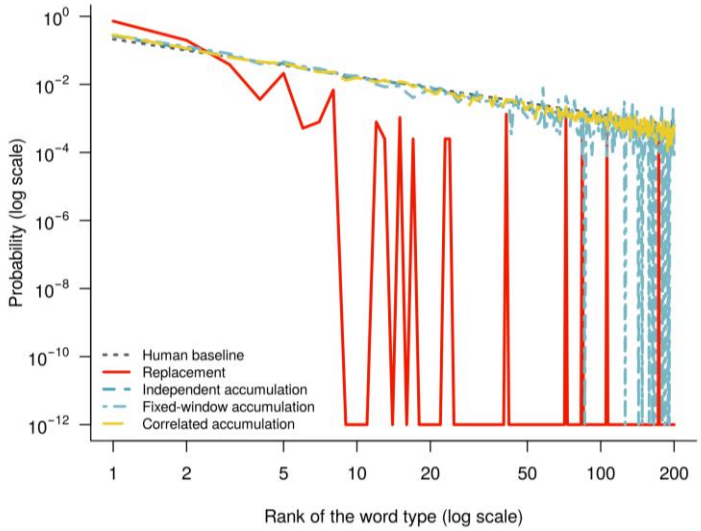
Bridge to the reduced-form state. This test reports several distinct distributional diagnostics associated with variety. Entropy, support, tail mass, and held-out loss need not move in identical order. Any scalar V_t therefore requires a declared mapping and validation criterion; the present experiment does not estimate a common direction of movement. Held-out cross-entropy also records predictive fit; it accompanies the distributional diagnostics without being silently identified with the bounded quality coordinate in Annex I.A.1.

Figure 1. Corpus diagnostics across generations



Note: Four panels: normalised entropy, held-out cross-entropy, vocabulary support and rare-tail mass for the four regimes. Lines are medians; shaded bands are the 10th–90th percentile range across histories.

Figure 2. Rank-frequency profiles in the final generation



Note: Lines: the four regimes and the human baseline.

The next exercise turns from which cases survive to what aggregation gains when many judgements inherit a common signal.

B. Test 2: Sharing an oracle

The statistical question in Sections 3–4 is whether an individually useful common source can erode the gains from combining different judgements. Annex I.A.2 supplies the exact diversity and variance results. Here a shared oracle enters a constructed estimation problem, allowing current error, prediction diversity, and residual correlation to be displayed together. The analytical saturation curve and the Monte Carlo outcomes are reported separately.

Each world contains a latent state. Human signals contain idiosyncratic noise, while a shared AI signal contributes a common error. Reliance varies from zero to one, and private-signal quality may erode with reliance. Tracked quantities are aggregate RMSE, within-crowd prediction diversity, residual correlation, and N_{eff}. The RMSE-minimising reliance may be zero, interior, or full delegation. Full reliance is a valid boundary result when the shared signal is sufficiently accurate. When the shared signal is perfect, residual correlation is not identified because every error is zero; the dashboard reports that the correction is unnecessary rather than treating the undefined correlation as a failure.

Benchmark calibration and results. The reported benchmark uses 40 agents and 300 independently generated worlds. Human private signals have noise standard deviation 1.00. The shared oracle has bias 0.30 and noise standard deviation 0.30. Reliance enters through convex blending, and the erosion parameter is set to zero so that this run removes the effort-erosion channel and examines the changing balance between idiosyncratic noise and a shared, biased signal. The dashboard evaluates the full reliance grid; Table 2 reports five selected points.

Table 2. Selected unpaired Monte Carlo points from the shared-oracle reliance grid

Reliance	RMSE	**Prediction diversity**	**Residual ρ**	N_{eff}
0.00	0.144	1.017	0.000	40.0
0.25	0.155	0.560	0.006	32.0
0.50	0.222	0.257	0.077	10.0
0.75	0.334	0.062	0.426	2.3
1.00	0.413	0.000	0.999	1.0

Note: Selected rows from an unpaired Monte Carlo run, seed 1. Residual ρ is the mean pairwise residual correlation held within [0, 0.999], the value used for the variance-equivalent sample size. Prediction diversity is the within-world sample variance of reports (divisor N − 1), that is N/(N − 1) times the diversity term in Page's identity. Parameters: agents=40; worlds=300; oracle bias=0.30; oracle noise=0.30; private noise=1.00; erosion=0.00; variant=blend. Worlds and primitive shocks were redrawn at each reliance value, so row differences include Monte Carlo noise.

Under the stated blend model, the population expectation is $\mathrm{MSE}(a) = \frac{(1-a)^2}{40} + 0.18a^2$. The continuous minimiser is $a^* = 0.121951$ and the 0.05-grid minimiser is $a = 0.10$, with RMSE 0.148492 versus 0.158114 at $a = 0$. The sample minimum of the reported run at $a = 0.15$ and its higher error at $a = 0.25$ than at a = 0 are properties of the unpaired Monte Carlo realisation, not of the stated expected-loss function; across 100 seeds the sample minimum lies between 0.05 and 0.20 in 96 runs.

Across the unpaired sample (Figure 3), prediction diversity falls from 1.017 to zero and the reported residual correlation rises from approximately zero to 0.999. The formula-based variance-equivalent sample-size diagnostic consequently falls from 40.0 to one while nominal N remains 40. These are descriptive sample outputs for an equal-weight mean; N_{eff} does not count independent voices or measure general information.

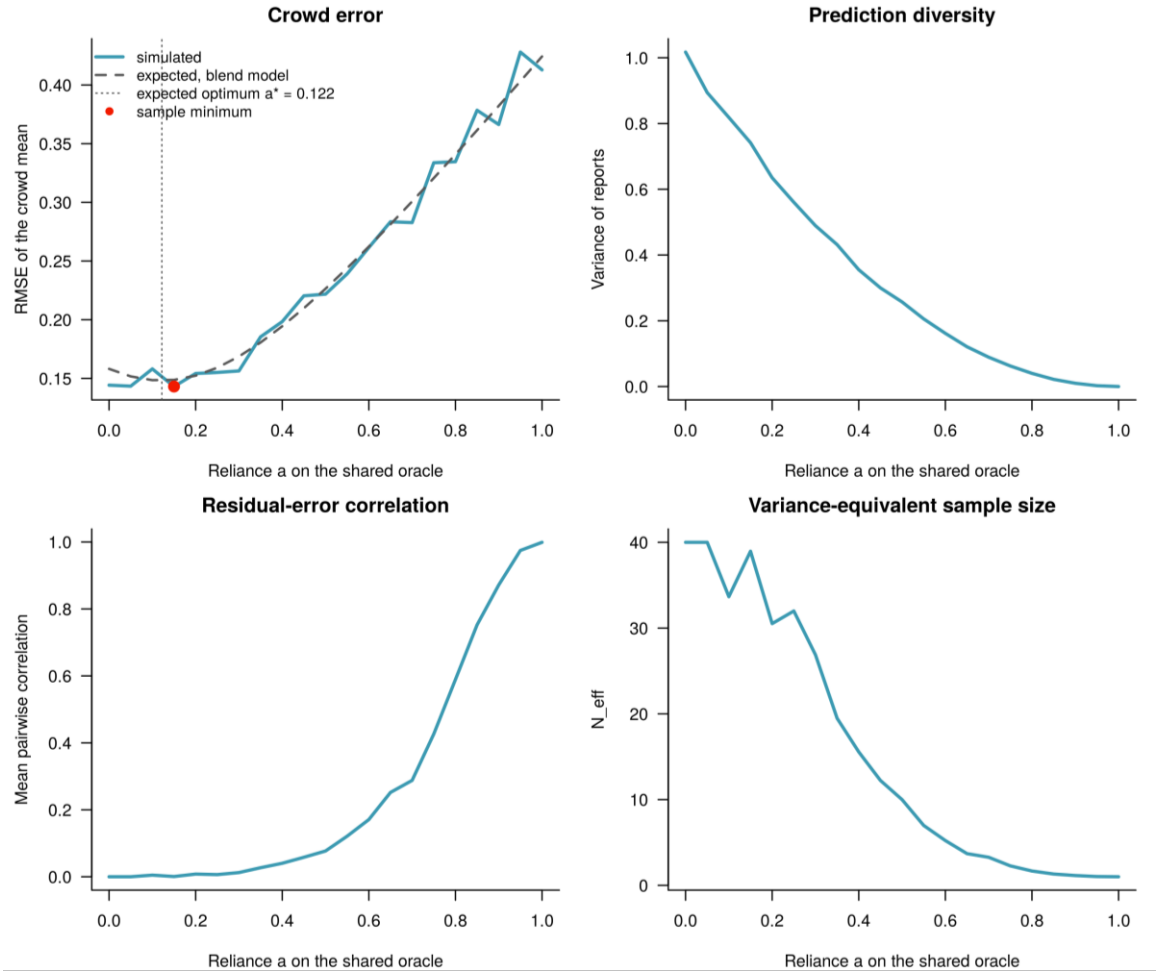
Figure 4 separates the analytical variance result from the benchmark simulation. Under equal weights, common residual variance, and equicorrelation, N_{eff} converges to $1/\rho$ for fixed $\rho > 0$. Correlations 0.05, 0.20, 0.50, and 0.80 imply asymptotic variance-equivalent limits 20, 5, 2, and 1.25 for that mean. This is a limit on variance reduction under the maintained estimator and error structure, not a general count of independent voices.

The exploratory parameter sweep in Figure 5 illustrates possible boundary outcomes but does not estimate their regions reliably. Grid points use separately drawn worlds, only ten replications, and tolerance-based labels for ‘ignore’ and ‘full’. Full delegation can be an exact boundary optimum when the shared signal is sufficiently accurate, but the displayed regime frequencies and jagged boundaries are provisional implementation outputs.

Sampling limits differ between the benchmark and the boundary sweep. The benchmark uses 300 worlds, while the boundary sweep uses ten replications at each bias-noise point with 20 agents and 60 worlds per replication. The sweep covers oracle noise from 0.05 to 1.5 in 16 steps and bias 0.05, 0.30 and 0.60, on a reliance grid with step 0.10; a replication is labelled ‘ignore’ when its sample minimum lies at reliance 0.1 or below and ‘full’ when it lies at 0.9 or above. Paired worlds, exact boundary definitions, and uncertainty on paired contrasts are required before interpreting the displayed regions.

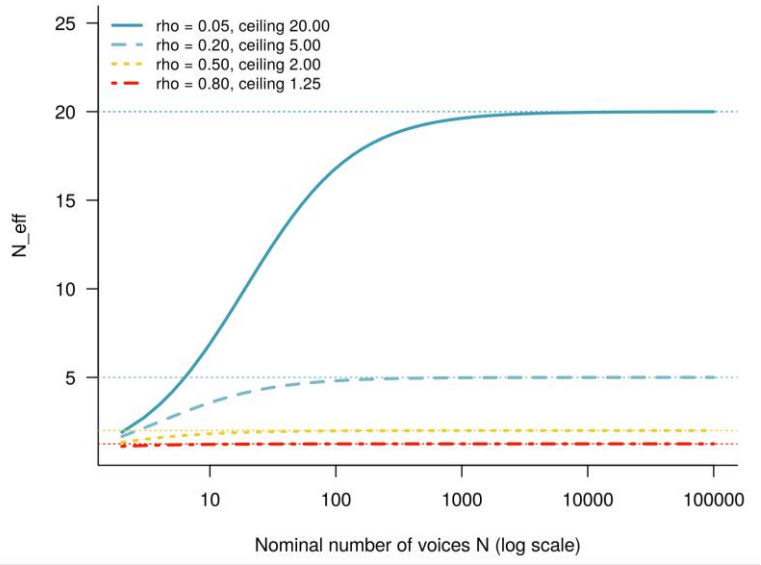
Bridge to the conditional recurrence. The benchmark imposes a reliance weight and a joint error structure while setting erosion to zero. It does not generate future human quality or endogenous reliance. The dashboard can separately impose a link from reliance to Q, but that feedback is a maintained law of motion rather than an output of the oracle experiment.

Figure 3. Shared-oracle reliance: Monte Carlo curves



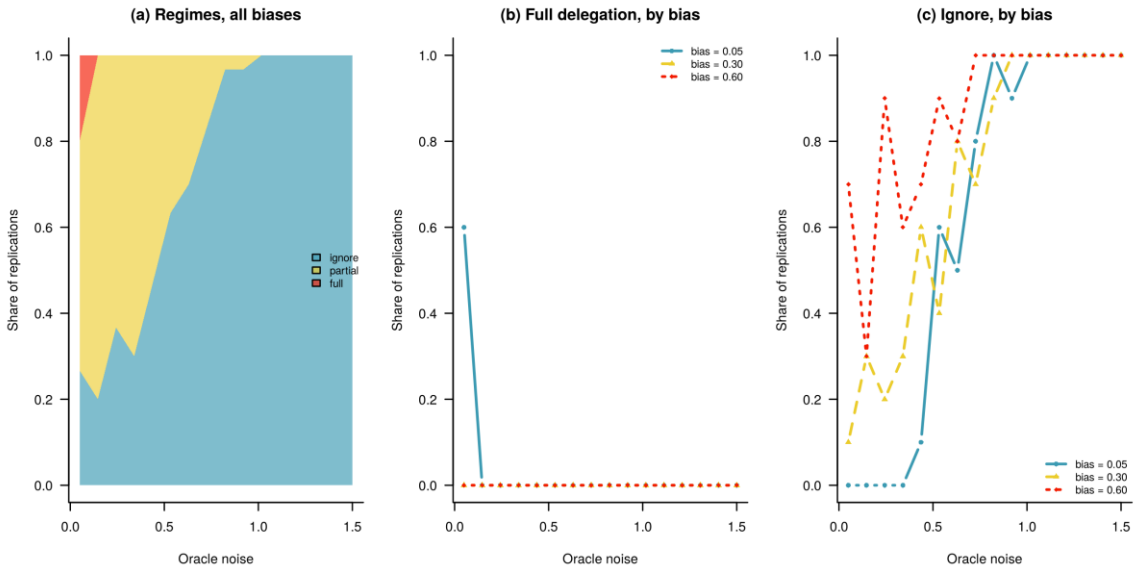
Note: Four panels against reliance a: crowd RMSE, prediction diversity, residual correlation and the variance-equivalent sample size. In the RMSE panel the dot marks the sample minimum, the dashed curve the expected error of the stated model and the dotted line its optimum.

Figure 4. Variance saturation (Proposition A2)



Note: Curves: variance-equivalent sample size against nominal N for fixed correlations ρ; horizontal lines are the limits 1/ρ.

Figure 5. Boundary sweep over oracle noise and bias



Note: Panel (a): regime shares averaged over bias; panel (b): probability of full delegation by bias; panel (c): probability of ignoring the oracle.

Test 2 shows that, under the stated blend model, a small oracle weight can lower expected MSE while greater common-source reliance reduces cross-sectional prediction diversity and the variance reduction obtained by averaging. This is a conditional estimator-specific result, not a general dissolution theorem for collective knowledge.

C. From a useful counterargument to population plurality

Section 6 gives the example of several users receiving the same strong counterargument. The following exercise asks what changes when the operator also uses local information to distribute alternatives across the population. It implements selected elements of Annex I.C, with value revealed after selection, and shows both a favourable calibration and one in which most alternatives fail validation.

Annex I.C defines Socratic behaviour at two levels. An exchange is individually Socratic when it introduces an underrepresented alternative that later proves relevant. A population-level Socratic operator must also avoid supplying the same alternative to every user. The dashboard compares a Mirror, a one-best Socratic operator guided by one common relevance signal, and a population-level operator that combines this signal with user-specific local information. Realised value is revealed only after the choice. The coded surprise is the log ratio of the peak model density to the model density at the candidate: the surprise measure of Annex I.C with the peak density as reference, before normalisation and capping at one. The surprise

thresholds 1.5 and 3.6 are on this log scale. The experiment therefore separates surprise at the moment of injection from downstream validation.

Baseline calibration. The reported baseline uses 200 users, a model centre of 0.35, concentration 12, a value frontier centred at 0.75 with width 0.15, a pool of 60 candidate alternatives, a surprise threshold of 1.5, and personalisation weight $p = 0.70$. Normalised cross-user dispersion lies in $[0,1]$. Table 3 reports the operating points generated in the same run as Figure 6. The model density is a Beta density with mean at the model centre and the stated concentration. Candidates are drawn uniformly on [0.01, 0.99]. Each user's value frontier is centred at frontier + 1.2 × width × u, with u uniform on [−1, 1], and the value of candidate x is $\exp(-(x - f)^2/(2\,\text{width}^2))$. The personalisation sweep uses steps of 0.05.

Table 3. Operator outcomes in the baseline quality-plurality calibration

Operator	Mean value	Far + valuable	Far + valueless	Normalised dispersion	Distinct alternatives
Mirror	0.045	0.00	0.00	0.000	1
Socratic: one best	0.736	0.77	0.23	0.000	1
Socratic: population	0.931	1.00	0.00	0.217	17

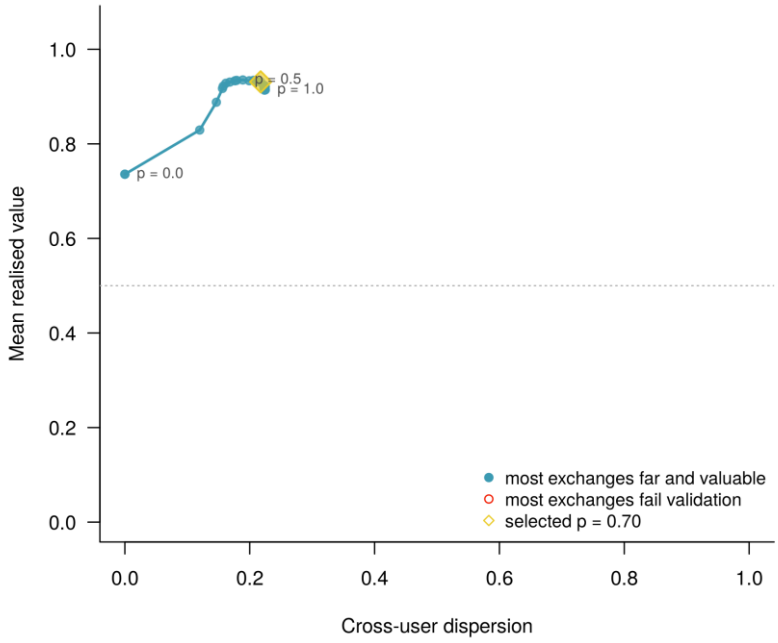
Note: All entries come from one dashboard run with seed 1, the run that also produces Figure 6. Near + valuable is zero for all three operators in this calibration. Mirror observations lie entirely in the near + valueless cell. An exchange counts as valuable when its realised value is at least 0.5 and as far when its surprise reaches the threshold. Operators score eligible candidates as 0.25 × surprise + 0.75 × relevance signal, both rescaled to [0, 1]; the public relevance signal carries noise with standard deviation 0.08 and the local signal noise with standard deviation 0.10.

The baseline lies in a favourable region. The Mirror supplies one concentrated answer with mean realised value 0.045. The one-best Socratic operator moves far from the model centre and raises mean value to 0.736, but its population signature remains concentrated in one alternative. Twenty-three and a half per cent of its exchanges are far but valueless after the hidden frontier is revealed. At $p = 0.70$, the population-level operator distributes 17 different alternatives, raises normalised dispersion to 0.217, and places every exchange in the far + valuable cell. Mean realised value increases to 0.931. Relative to one-best, current value rises by 0.195, the validated share rises by 23.5 percentage points, and plurality expands from one to 17 alternatives. Across 100 seeds the population-level operator exceeds one-best Socrates in mean value in every world.

Quality and coordinate dispersion rise together in this generated run because personalisation exploits local signals produced by the same favourable one-dimensional value

generator. Figure 6 shows the resulting high-value plateau: personalisation moves the population from one common counterargument towards multiple locally informed alternatives. Its interpretation belongs to this calibration; a general causal claim about personalisation would also have to confront adverse and null calibrations, and higher coded value does not establish a welfare gain.

Figure 6. Quality-plurality frontier, baseline calibration



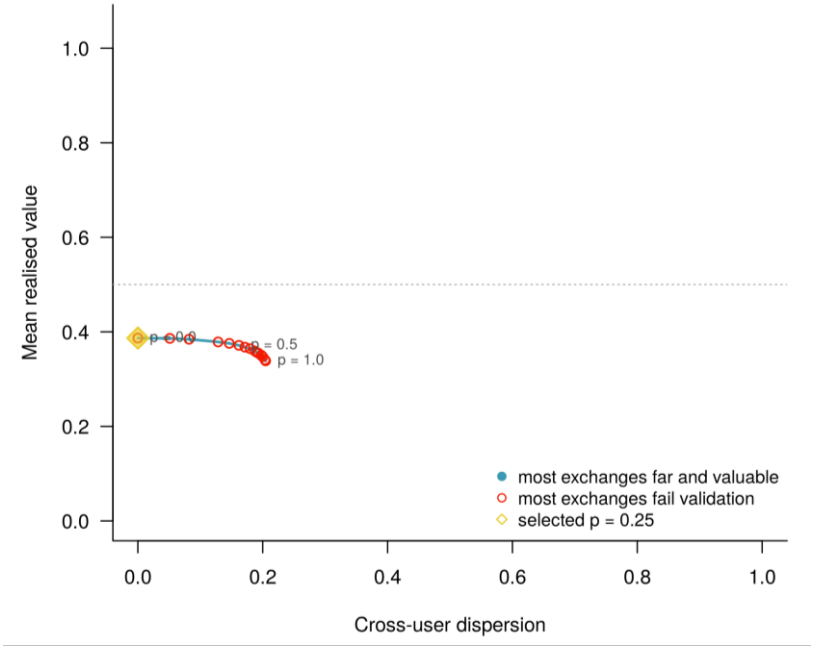
Note: The curve traces mean realised value against cross-user dispersion as personalisation p rises from 0 to 1; the marked point is p = 0.70.

Contrasting calibration: a validation-threshold region. Figure 7 changes the environment rather than the plotting convention. It uses 300 users, concentration 15, a frontier at 0.45 with width 0.19, a pool of 30 candidates, and a surprise threshold of 3.6. At the selected point $p = 0.25$, mean realised value is 0.387, normalised dispersion is zero, and only 33.7 per cent of exchanges fall in the far + valuable cell: personalisation has not yet produced plurality, and most deviations fail validation. In the reported realisation stronger personalisation creates plurality without validated value. At full personalisation dispersion reaches 0.205, the far + valuable share stays at 33.3 per cent, and mean value falls to 0.339. Across 100 seeds, full personalisation raises mean value above its level at p = 0.25 in 84 worlds, and a majority of far and valuable exchanges appears in 31.

The comparison reveals two shapes of the same mechanism. In the baseline, local information complements the common signal almost immediately, so current value

and population plurality improve together. In the validation-threshold calibration, plurality can grow while validated value does not, and whether value follows depends on the realised world. Dispersion alone therefore cannot be treated as the social output. The relevant flow is dispersion attached to contributions that survive delayed validation.

Figure 7. Quality-plurality frontier, validation-threshold calibration



Note: The curve traces mean realised value against cross-user dispersion as personalisation p rises from 0 to 1; the marked point is p = 0.25.

Bridge from operator dispersion to formal measurement. The experiment reports a one-dimensional cross-user dispersion statistic. It does not estimate the validated contribution law, the residual-error kernel, the source-similarity kernel, or the cross-component mixture integrals.

Test 3 examines coded submission and validation outputs. Corpus replenishment through the exact mixture transition remains a separate implementation requirement.

D. Test 3: Pricing friction and validation

The institutional argument in Section 7 distinguishes paying for surprise from paying for a contribution that survives scrutiny. This exercise compares no renewal reward, immediate surprise payment, and delayed central or plural validation. Its formal counterparts are the choices and accounts in Annex I.D and the verification problem in Annex I.E. The outputs show what the coded rules purchase in selected populations, not whether the complete institution is feasible.

The code compares four local rules with heterogeneous privacy costs and a simplified privacy-based sharing decision, but users do not face the participation outside option specified in Annex I.D.1. Sharing uses mode-specific pilot beliefs rather than the full contribution-conditioned payment law; central and plural architectures generate separate pilot beliefs. Energy requirements are drawn independently for the two modes, so the Socratic mode uses less energy than the Mirror for about a quarter of users (24 per cent in the reported run); the run therefore does not impose the positive gap Δe of Annex I.D.1a on every user. The reported calibration records coded Socratic production, payments, rewarded positions, and a one-period deposit-only balance. It does not implement provenance, appeal, strategic validation, unresolved claims, operating costs, or the reserve envelope. Its incentive comparisons consequently leave participation, privacy, welfare, and intertemporal feasibility to the fuller institutional assessment.

Reported calibration. The current point uses 400 users, median friction 0.35, median privacy cost 0.12, parameter vector $b_M = 1.00$, $b_S = 0.90$, $e_M = 1.0$, $e_S = 1.6$, $d = 0.30$, energy price 0.050, full rebate of the deposit, coded royalty $r^{\text{code}} = 0.30$, seven verifiers, verifier spread 0.80, and quorum share 0.35. The coded royalty corresponds to the formal prize ϖ only in this single-period implementation. Architecture-specific validation beliefs are expected quorum probabilities estimated in an unpriced pilot pass. Friction, privacy cost and the energy requirements of the two modes are drawn per user from lognormal distributions; the values above are their medians. The reported run uses seed 1. The standard deviations of the logarithms are 0.6 for friction, 0.7 for privacy cost and 0.5 for the energy requirement of each mode; a paid contribution counts as a false reward when its realised value is below 0.30. Verifiers sit at evenly spaced positions around the model centre, spread × [−0.45, 0.45], and all at the centre under central validation. A verifier accepts a contribution with probability logistic(9 × (value × (0.25 + 0.75 × fit) − 0.42)), where fit is a Gaussian function of the distance to its position with width max(0.16, 1.6 × frontier width). Votes are independent given the contribution, so the coded joint vote law is a product law; acceptance requires ⌈quorum share × V⌉ votes. Under the immediate subsidy users choose by 0.60 × surprise + 0.40 × noise.

Three institutional signatures (Table 4). The immediate surprise subsidy induces Socratic mode in 41 per cent of the coded users and generates the largest coded flow, $F^{\text{code}} = 3.040$. This coded volume is obtained with a false-reward share of 49.1 per cent. The mean realised value of paid contributions is 0.399, rewarded dispersion is 0.028, and only eight positions receive payment. In Figure 8 this is the point with the highest false-reward share and the largest

flow. Surprise payment therefore buys a large quantity of model-distant output, while almost half of the rewarded contributions fail the later value test. The coded flow cannot serve as a sufficient performance statistic.

Table 4. Institutional outcomes in the pricing-friction calibration

Institution	Socratic share	Rewarded share	F^{code}	Mean value paid	False rewards	Dispersion / positions	Budget / user
No renewal	0.000	0.000	0.000	0.000	0.000	0.000 / 0	0.000
Immediate surprise	0.412	0.412	3.040	0.399	0.491	0.028 / 8	0.052
Delayed central	0.010	0.000	0.000	0.000	0.000	0.000 / 0	0.300
Delayed plural	0.190	0.133	0.636	0.945	0.000	0.229 / 15	0.221

Note: Seed 1. In the implementation, F^{code} denotes the population mean of the coded, unnormalised structured-novelty score, not a sum. It is distinct from the formal expected flow $F_t^g \in [0,1]$. The coded flow is a local mechanism output and should be read together with false rewards, mean realised value, rewarded dispersion, and the one-period cash balance. Dispersion is normalised to the interval [0,1].

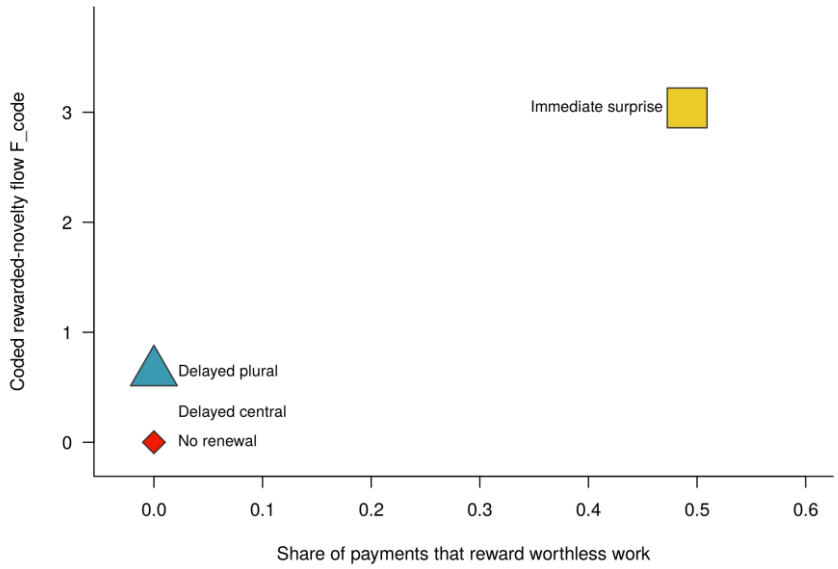
Central delayed validation is almost inactive; in Figure 8 its point coincides with no renewal at the origin. In the reported run 1.0 per cent of users choose the Socratic mode and no contribution is paid, so the coded flow and the mean value of paid contributions are zero. Across 40 seeds the central verifier pays a median of one position, and in 9 of them it pays no one. The criterion-profile correlation and the legacy linear variance-design diagnostic both equal one, but neither number establishes verifier precision, replenishment, or a quorum operating characteristic.

In Figure 9 the desired region combines high realised value with dispersion across many rewarded positions. Plural delayed validation reports Socratic share 0.190, rewarded share 0.133, mean realised value 0.945, rewarded dispersion 0.229, and one-period deposit-minus-payout balance 0.221 per user in this run. The criterion-profile correlation is 0.104 and the legacy linear variance-design diagnostic equals 4.311, but that number is not an effective quorum count and does not determine k-of-V acceptance. Because the pilot beliefs, submitted pool, criteria, and vote law are generated within the same calibration, these values do not establish that functional plurality has causal economic force.

The architecture sweep varies verifier spread and quorum together. Greater spread changes latent criterion profiles and the legacy linear variance-design diagnostic, but pairwise profile correlation does not determine the k-of-V acceptance probability. The plotted changes

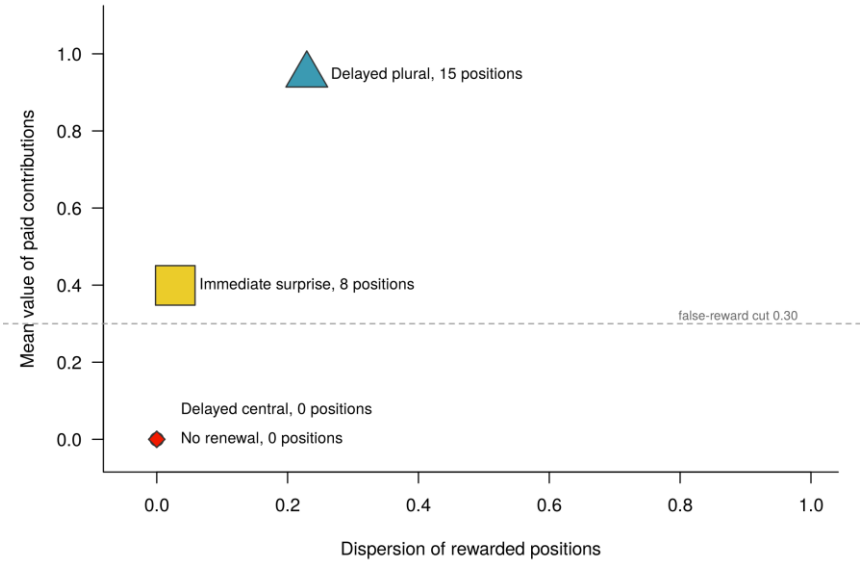
in validation belief and realised flow are compound implementation outcomes; verifier accuracy, joint vote dependence, criterion coverage, and quorum must be varied and reported separately.

Figure 8. Institutional operating points



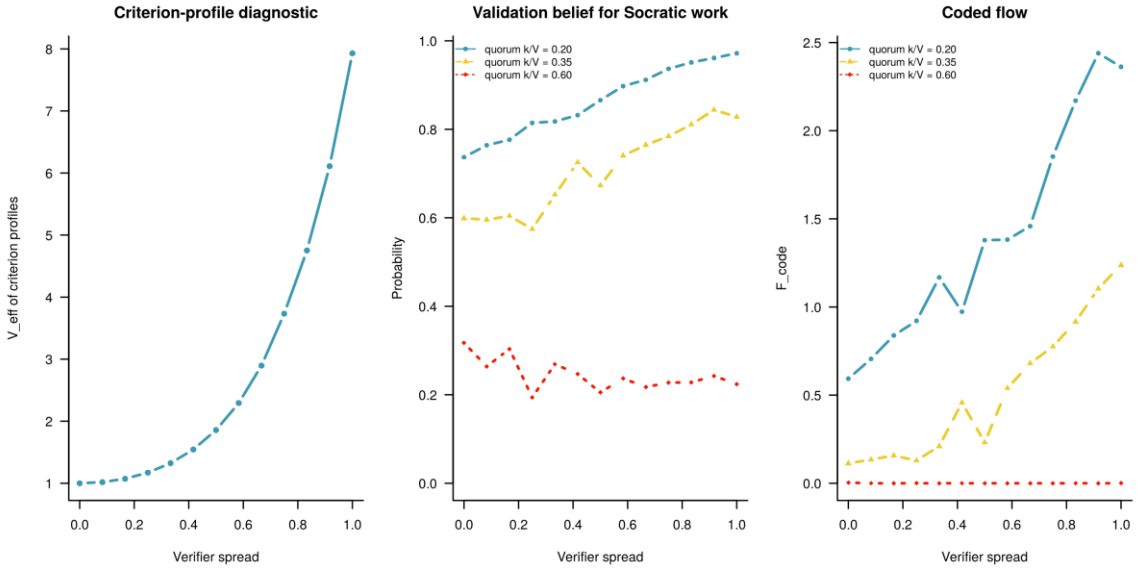
Note: Horizontal axis: share of payments for worthless work; vertical axis: coded rewarded-novelty flow; point size: Socratic adoption.

Figure 9. Quality and plurality of rewards



Note: Horizontal axis: dispersion of rewarded positions; vertical axis: mean realised value of paid contributions; point size: rewarded share; dashed line: false-reward cut 0.30.

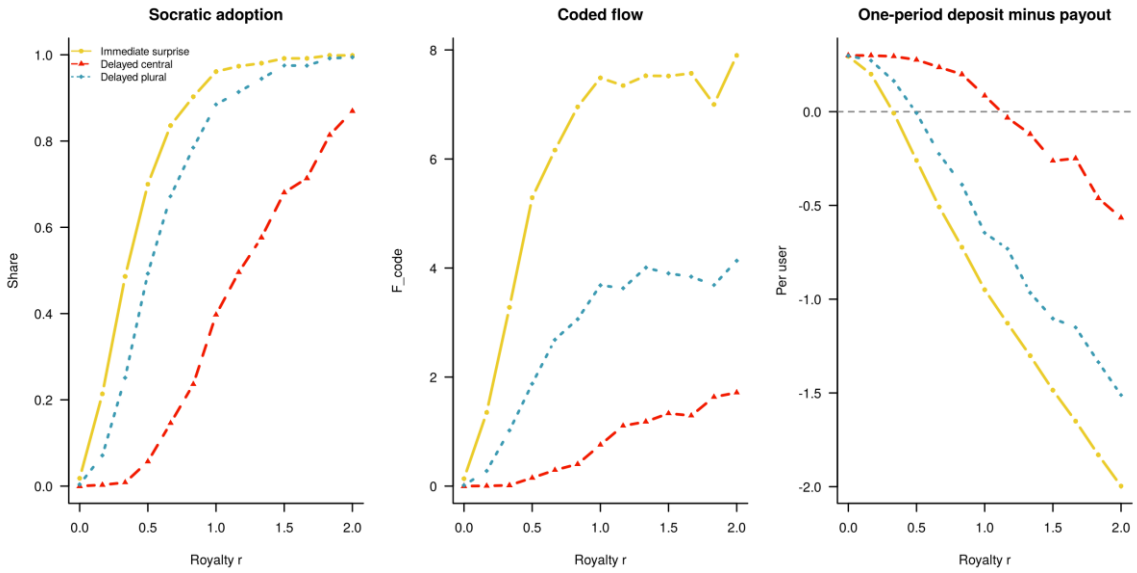
Figure 10. Verification-architecture sweep



Note: Left panel: legacy linear variance-design diagnostic of criterion profiles (V_eff on the axis); middle: validation belief for Socratic work; right: coded flow. Lines: quorum shares k/V.

Royalty sweep. Increasing r^{code} changes coded Socratic adoption, realised flow, and the one-period deposit-minus-payout balance across independently redrawn populations. A non-negative one-period balance neither verifies the formal reserve inequality nor identifies a feasible royalty interval when rewards vest later. That test requires fixed worlds, cohort liabilities, operating and verification costs, and the no-borrowing reserve rule.

Figure 11. Royalty sweep



Note: Panels: Socratic adoption, coded flow and one-period deposit-minus-payout balance against the royalty r, for three institutions.

Boundary of what Test 3 establishes. The experiment identifies candidate margins for a fuller design: operator choice, realised value, false rewards, and rewarded dispersion. Participation and cohort solvency require the additional implementation specified in Annex I.D. The reported calibration does not estimate them on a common causal or equilibrium basis, establish corpus replenishment, or demonstrate financial feasibility.

E. Conditional trajectories under constant inputs

Section 5 asks what should be preserved when quality, variety, and error dependence do not move together. To make that tension concrete, the final exercise applies the recurrence in Annex I.B with fixed inputs. It traces the coordinates separately, rather than using an improvement in one diagnostic to declare that renewal has succeeded.

The final exercise applies the projected recurrence with supplied inputs and zero shocks. It checks arithmetic under one maintained sign specification.

Reported calibration and stationary state. The trajectory begins at $(Q_0, V_0, \rho_0) = (0.50, 0.50, 0.30)$ and is simulated for 40 periods. Quality uses $\mu_Q = 0.06$ and $\gamma_Q = 0.08$. Variety uses $\mu_V = 0.05$, $\gamma_h = 0.08$, $\gamma_F = 0.05$, $\lambda_V = 0.06$, fresh inflow $h = 0.50$, replacement $\xi = 0.50$, and validated flow $F = 0.20$. Correlation uses $\mu_\rho = 0.08$, $\gamma_\rho = 0.08$, $\lambda_\rho = 0.12$, reliance $a = 0.50$, and population dispersion $\Sigma = 0.30$. Conditional on these supplied flows, the interior stationary state is $Q^* = 0.667$, $V^* = 0.400$, and $\rho^* = 0.050$. The simulated path approaches this state monotonically. For this calibration, the learning input is set to one minus reliance, and the source-similarity input is set equal to reliance. These imposed links produce the stated stationary values; they are not estimates from the preceding experiments.

A mixed numerical trajectory (Table 5; Figures 12–13). Under the supplied coefficients, quality rises, the coded residual-error dependence coordinate falls, and variety declines. The constructed D and variance-design statistics rise by construction. Without an estimated loss or welfare functional, the model cannot conclude that gains in quality or lower dependence outweigh the loss of variety.

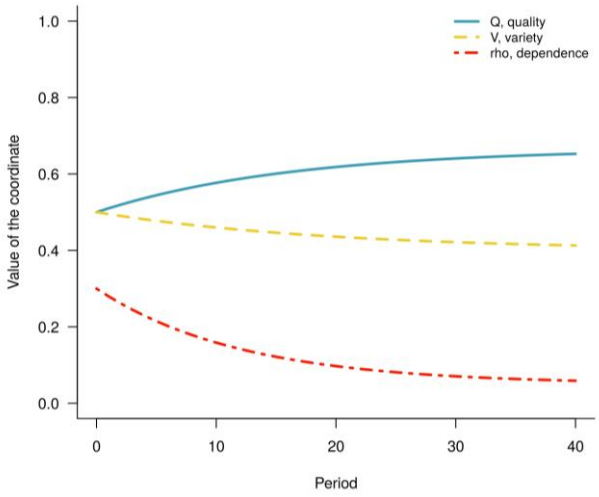
Why the secondary diagnostics separate. Their difference follows mechanically from different transformations of the same supplied state. It provides a sensitivity comparison, not independent evidence of adaptive capacity, effective complexity, or welfare. Neither statistic is used to support the paper's central contribution.

Table 5. State and diagnostics along the reduced-form trajectory

Period	Q	V	ρ^{err}	D	C^*	N_{eff}
0	0.500	0.500	0.300	0.175	0.350	3.0
8	0.565	0.466	0.178	0.217	0.462	4.6
16	0.605	0.444	0.116	0.237	0.528	6.2
24	0.629	0.429	0.084	0.247	0.565	7.7
32	0.644	0.419	0.067	0.252	0.585	8.8
40	0.653	0.413	0.059	0.254	0.596	9.4

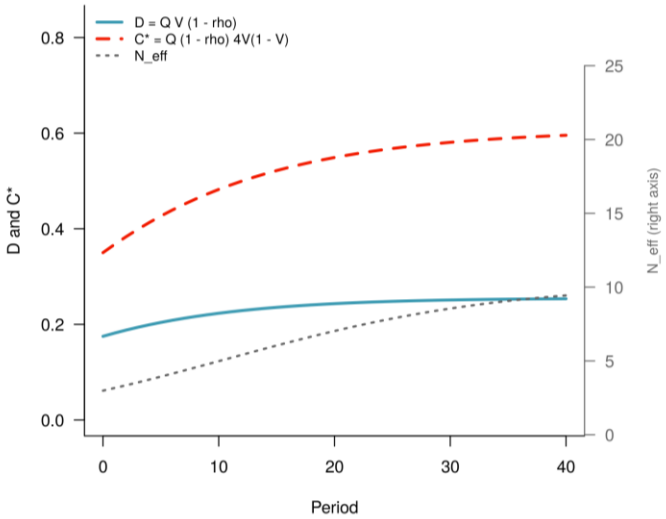
Note: $D_t := Q_t V_t(1 - \rho_t^{err})$ *and the legacy dashboard complexity statistic are constructed secondary diagnostics. The variance-equivalent sample-size statistic uses nominal N=20 and the displayed residual-error dependence under an equal-weight equicorrelation design; it is not a count of independent voices. The figures retain the short legacy label rho for this coded coordinate. The legacy statistic is C* = Q(1 − ρ)·4V(1 − V).*

Figure 12. Conditional paths of the three coordinates



Note: Lines: quality Q, variety V and residual-error dependence (legend label rho).

Figure 13. Secondary constructed diagnostics



Note: Left axis: D and C; right axis: variance-equivalent sample size N_eff (dotted).*

Boundary of the conditional recurrence. Reliance, inflow, replacement, validated flow, source similarity, and dispersion are supplied rather than jointly estimated or passed from earlier modules.

Bridge to formal measurement. The exact mixture identities construct validated contribution laws, update the corpus distribution, and evaluate quality, variety, residual-error dependence, and source similarity as separate declared functionals. Their numerical implementation would provide the missing link between the earlier contribution exercises and these state trajectories. The present annex therefore shows connected questions and separately computed illustrations, rather than a single fitted simulation running from dialogue to renewal.

Further formal development depends on preregistering and validating the measurement protocol, estimating the conditional signs, and comparing candidate operators and validator architectures on fixed tasks. If these empirical objects support it, a separate mechanism-design paper could then add strategic participation, manipulation, and welfare structure.

References

Acemoglu, D., 2025. The simple macroeconomics of AI. Economic Policy, 40(121), 13–58. https://doi.org/10.1093/epolic/eiae042

Acemoglu, D., Kong, D., Ozdaglar, A., 2026. AI, human cognition and knowledge collapse (NBER Working Paper No. 34910). National Bureau of Economic Research. https://doi.org/10.3386/w34910

Akerlof, G.A., 1970. The market for 'lemons': Quality uncertainty and the market mechanism. The Quarterly Journal of Economics, 84(3), 488–500. https://doi.org/10.2307/1879431

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A.I., Babaei, H., LeJeune, D., Siahkoohi, A., Baraniuk, R.G., 2024. Self-consuming generative models go MAD, in: The Twelfth International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2024/hash/ebc042e767de551803ccfcc45e2454f5-Abstract-Conference.html

Allen, R.C., 1982. The efficiency and distributional consequences of eighteenth-century enclosures. The Economic Journal, 92(368), 937–953. https://doi.org/10.2307/2232676

Arrow, K.J., 1962. Economic welfare and the allocation of resources for invention, in: Nelson, R.R. (Ed.), The rate and direction of inventive activity: Economic and social factors. Princeton University Press, pp. 609–626. https://doi.org/10.1515/9781400879762-024

Arrow, K.J., Fisher, A.C., 1974. Environmental preservation, uncertainty, and irreversibility. The Quarterly Journal of Economics, 88(2), 312–319. https://doi.org/10.2307/1883074

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., Mariman, R., 2025. Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

Baumann, F., Akçay, E., Plotkin, J.B., 2026. Generative artificial intelligence reduces social welfare through model collapse. arXiv. https://doi.org/10.48550/arXiv.2604.21853

Beff Jezos, bayes, 2022. Notes on e/acc principles and tenets: A physics-first view of the principles underlying effective accelerationism. Beff's Newsletter, July 10. https://beff.substack.com/p/notes-on-eacc-principles-and-tenets

Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, pp. 610–623. https://doi.org/10.1145/3442188.3445922

Bennett, C.H., 1982. The thermodynamics of computation—A review. International Journal of Theoretical Physics, 21, 905–940. https://doi.org/10.1007/BF02084158

Bestvater, S., Smith, A., TerBush, C., Baronavski, C., Chavda, J., 2026. How much of the internet is written with AI? Pew Research Center, August 20. https://www.pewresearch.org/data-labs/2026/08/20/how-much-of-the-internet-is-written-with-ai/ (accessed 18.10.26)

Bloom, N., Jones, C.I., Van Reenen, J., Webb, M., 2020. Are ideas getting harder to find? American Economic Review, 110(4), 1104–1144. https://doi.org/10.1257/aer.20180338

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A.S., Creel, K., Davis, J.Q., Demszky, D., …, Liang, P., 2021. On the opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Boyle, J., 2003. The second enclosure movement and the construction of the public domain. Law and Contemporary Problems, 66(1–2), 33–74.

Brynjolfsson, E., Li, D., Raymond, L., 2025. Generative AI at work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

Calvano, E., Haghtalab, N., Vitercik, E., Zhao, E., 2024. Algorithmic content selection and the impact of user disengagement. arXiv. https://doi.org/10.48550/arXiv.2410.13108

Campbell, M., Hoane, A.J., Jr., Hsu, F.-h., 2002. Deep Blue. Artificial Intelligence, 134(1–2), 57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Cao, W., 2026. Position: Generative models erode temporal learning through market selection, in: Proceedings of the 43rd International Conference on Machine Learning (ICML 2026). Proceedings of Machine Learning Research. https://doi.org/10.48550/arXiv.2606.06572

Chen, N., Tong, Y., Yang, Y., He, Y., Zhang, X., Zou, Q., Wang, Q., He, B., 2026. Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation, in: Findings of the Association for Computational Linguistics: ACL 2026. Association for Computational Linguistics, pp. 251–306. https://doi.org/10.18653/v1/2026.findings-acl.13

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L.K., Aitchison, M., Orseau, L., Hutter, M., Veness, J., 2024. Language modeling is compression, in: The Twelfth International Conference on Learning Representations. https://arxiv.org/abs/2309.10668

del Rio-Chanona, R.M., Laurentsyeva, N., Wachs, J., 2024. Large language models reduce public knowledge sharing on online Q&A platforms. PNAS Nexus, 3(9), pgae400. https://doi.org/10.1093/pnasnexus/pgae400

Dohmatob, E., Feng, Y., Yang, P., Charton, F., Kempe, J., 2024. A tale of tails: Model collapse as a change of scaling laws, in: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, Vol. 235, pp. 11165–11197. https://proceedings.mlr.press/v235/dohmatob24b.html

Doshi, A.R., Hauser, O.P., 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Science Advances, 10(28), eadn5290. https://doi.org/10.1126/sciadv.adn5290

Eloundou, T., Manning, S., Mishkin, P., Rock, D., 2024. GPTs are GPTs: Labor market impact potential of LLMs. Science, 384(6702), 1306–1308. https://doi.org/10.1126/science.adj0998

Emerson, R.W., 1841. Self-reliance, in: Essays: First series. James Munroe and Company.

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., Ruiz, F.J.R., Schrittwieser, J., Swirszcz, G., Silver, D., Hassabis, D., Kohli, P., 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. Nature, 610, 47–53. https://doi.org/10.1038/s41586-022-05172-4

Feng, Y., Dohmatob, E., Yang, P., Charton, F., Kempe, J., 2025. Beyond model collapse: Scaling up with synthesized data requires verification, in: The Thirteenth International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2025/hash/df2d62b96a4003203450cf89cd338bb7-Abstract-Conference.html

Fromm, E., 1941. Escape from freedom. Farrar & Rinehart.

Fromm, E., 1973. The anatomy of human destructiveness. Holt, Rinehart and Winston.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D.A., Yang, D., Donoho, D.L., Koyejo, S., 2024. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data, in: Proceedings of the Conference on Language Modeling (COLM 2024). https://openreview.net/pdf?id=5B2K4LRgmz

Gigerenzer, G., 2008. Rationality for mortals: How people cope with uncertainty. Oxford University Press. https://doi.org/10.1093/oso/9780195328981.001.0001

Gigerenzer, G., Selten, R. (Eds.), 2001. Bounded rationality: The adaptive toolbox. MIT Press.

Glickman, M., Sharot, T., 2025. How human–AI feedback loops alter human perceptual, emotional and social judgements. Nature Human Behaviour, 9, 345–359. https://doi.org/10.1038/s41562-024-02077-2

Granovetter, M., 1985. Economic action and social structure: The problem of embeddedness. American Journal of Sociology, 91(3), 481–510. https://doi.org/10.1086/228311

Gunderson, L.H., Holling, C.S. (Eds.), 2002. Panarchy: Understanding transformations in human and natural systems. Island Press. https://doi.org/10.2307/jj.41003716

Hayek, F.A., 1945. The use of knowledge in society. American Economic Review, 35(4), 519–530.

Holling, C.S., 1973. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics, 4, 1–23. https://doi.org/10.1146/annurev.es.04.110173.000245

Holmström, B., 1979. Moral hazard and observability. The Bell Journal of Economics, 10(1), 74–91. https://doi.org/10.2307/3003320

Hong, L., Page, S.E., 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proceedings of the National Academy of Sciences, 101(46), 16385–16389. https://doi.org/10.1073/pnas.0403723101

Hutchins, E., 1995. Cognition in the wild. MIT Press.

Jones, C.I., Tonetti, C., 2020. Nonrivalry and the economics of data. American Economic Review, 110(9), 2819–2858. https://doi.org/10.1257/aer.20191330

Kahneman, D., 1973. Attention and effort. Prentice-Hall.

Kahneman, D., 2011. Thinking, fast and slow. Farrar, Straus and Giroux.

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica, 47(2), 263–292. https://doi.org/10.2307/1914185

Kallel, M., El Louadi, M., 2026. Cognitive commons in the age of generative intelligence: A heterodox appraisal of the knowledge erosion hypothesis. arXiv. https://doi.org/10.48550/arXiv.2607.13272

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization, in: Proceedings of ICNN'95—International Conference on Neural Networks. IEEE, Vol. 4, pp. 1942–1948. https://doi.org/10.1109/ICNN.1995.488968

Kim, E.M., Garg, A., Peng, K., Garg, N., 2025. Correlated errors in large language models, in: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, Vol. 267, pp. 30038–30066. https://proceedings.mlr.press/v267/kim25e.html

Kleinberg, J., Raghavan, M., 2021. Algorithmic monoculture and social welfare. Proceedings of the National Academy of Sciences, 118(22), e2018340118. https://doi.org/10.1073/pnas.2018340118

Kopczewski, T., 2026a. Know Thyself: A methodological manifesto for teaching microeconomics through epistemic provocation (Working Papers 2026-14). Faculty of Economic Sciences, University of Warsaw. https://ideas.repec.org/p/war/wpaper/2026-14.html

Kopczewski, T., 2026b. Know Thyself Prompt Library: Replication Labs, Ad Hoc Research Dashboards, and ABM Tools [Working repository]. https://github.com/tomvar/know-thyself-prompt-library

Kopczewski, T., Lisicki, J., Potocki, T., 2026. From didactics to discovery: 'Entropy Lab' as a proof-of-concept for modelling uncertainty. Nonlinear Dynamics, Psychology, and Life Sciences, 30(3), 371–403.

Kopczewski, T., Potocki, T., Jaworski, N., Lisicki, J., Okhrimenko, K., Wójcik, A., n.d. When does collective learning become collective capture? Social learning models, algorithmic trolls, and generative AI [Manuscript submitted for publication].

Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.-H., Beresnitzky, A.V., Braunstein, I., Maes, P., 2025. Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. arXiv. https://doi.org/10.48550/arXiv.2506.08872

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25. Curran Associates, pp. 1097–1105.

Laitinen-Fredriksson Lundström-Imanov, G.O.Y., 2026. The economics of model collapse: Equilibrium, welfare, and optimal provenance subsidies in synthetic data markets. arXiv. https://doi.org/10.48550/arXiv.2605.20279

Landauer, R., 1961. Irreversibility and heat generation in the computing process. IBM Journal of Research and Development, 5(3), 183–191. https://doi.org/10.1147/rd.53.0183

Lorenz, J., Rauhut, H., Schweitzer, F., Helbing, D., 2011. How social influence can undermine the wisdom of crowd effect. Proceedings of the National Academy of Sciences, 108(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

Luccioni, A.S., Jernite, Y., Strubell, E., 2024. Power hungry processing: Watts driving the cost of AI deployment?, in: Proceedings of the 2024 ACM Conference on Fairness,

Accountability, and Transparency. Association for Computing Machinery, pp. 85–99. https://doi.org/10.1145/3630106.3658542

March, J.G., 1991. Exploration and exploitation in organizational learning. Organization Science, 2(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Mas-Colell, A., Whinston, M.D., Green, J.R., 1995. Microeconomic theory. Oxford University Press.

McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E., 2006. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AI Magazine, 27(4), 12–14. https://doi.org/10.1609/aimag.v27i4.1904 (Original work published 1955)

Mercier, H., Sperber, D., 2017. The enigma of reason. Harvard University Press. https://doi.org/10.4159/9780674977860

Nadella, S., 2026. The reverse information paradox. sn scratchpad, July 12. https://snscratchpad.com/posts/reverse-information-paradox/ (accessed 18.10.26)

Ostrom, E., 1990. Governing the commons: The evolution of institutions for collective action. Cambridge University Press. https://doi.org/10.1017/CBO9780511807763

Padmakumar, V., He, H., 2024. Does writing with language models reduce content diversity?, in: The Twelfth International Conference on Learning Representations. https://proceedings.iclr.cc/paper_files/paper/2024/hash/02dec8877fb7c6aa9a79f81661baca7c-Abstract-Conference.html

Page, S.E., 2007. The difference: How the power of diversity creates better groups, firms, schools, and societies. Princeton University Press. https://doi.org/10.2307/j.ctt7sp9c

Savage, L.J., 1954. The foundations of statistics. Wiley.

Sen, A.K., 1977. Rational fools: A critique of the behavioral foundations of economic theory. Philosophy & Public Affairs, 6(4), 317–344.

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal, 27(3), 379–423; 27(4), 623–656.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y., 2023. The curse of recursion: Training on generated data makes models forget. arXiv. https://doi.org/10.48550/arXiv.2305.17493

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., Gal, Y., 2024. AI models collapse when trained on recursively generated data. Nature, 631, 755–759. https://doi.org/10.1038/s41586-024-07566-y

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., 2016. Mastering the game of Go with deep neural networks and tree search. Nature, 529, 484–489. https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. Nature, 550, 354–359. https://doi.org/10.1038/nature24270

Simon, H.A., 1955. A behavioral model of rational choice. The Quarterly Journal of Economics, 69(1), 99–118. https://doi.org/10.2307/1884852

Simon, H.A., 1971. Designing organizations for an information-rich world, in: Greenberger, M. (Ed.), Computers, communications, and the public interest. Johns Hopkins Press, pp. 37–72.

Simon, H.A., 1990. Invariants of human behavior. Annual Review of Psychology, 41, 1–20. https://doi.org/10.1146/annurev.ps.41.020190.000245

Sims, C.A., 2003. Implications of rational inattention. Journal of Monetary Economics, 50(3), 665–690. https://doi.org/10.1016/S0304-3932(03)00029-1

Smith, V.L., 2003. Constructivist and ecological rationality in economics. American Economic Review, 93(3), 465–508. https://doi.org/10.1257/000282803322156954

Smith, V.L., 2008. Rationality in economics: Constructivist and ecological forms. Cambridge University Press. https://doi.org/10.1017/CBO9780511754364

Sparrow, B., Liu, J., Wegner, D.M., 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. Science, 333(6043), 776–778. https://doi.org/10.1126/science.1207745

Surowiecki, J., 2004. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Doubleday.

Szilárd, L., 1964. On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. Behavioral Science, 9(4), 301–310. https://doi.org/10.1002/bs.3830090402 (Original work published 1929)

Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving decisions about health, wealth, and happiness. Yale University Press. https://yalebooks.co.uk/book/9780300146813/nudge/

Turing, A.M., 1950. Computing machinery and intelligence. Mind, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

U.S. Copyright Office, 2025. Copyright and artificial intelligence, Part 3: Generative AI training [Pre-publication report]. https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

Weisbrod, B.A., 1964. Collective-consumption services of individual-consumption goods. The Quarterly Journal of Economics, 78(3), 471–477. https://doi.org/10.2307/1879478

Wikimedia Foundation, 2025. New user trends on Wikipedia. Diff, October 17. https://diff.wikimedia.org/2025/10/17/new-user-trends-on-wikipedia/ (accessed 18.10.26)

Wu, T., 2016. The attention merchants: The epic scramble to get inside our heads. Knopf.

Zuboff, S., 2019. The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.



University of Warsaw
Faculty of Economic Sciences
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl
ISSN 2957-0506