Perturbatics:
An Agentoscope for Reading Agency

PIATRA . INSTITUTE

July 2026

Abstract

At rest, a system reveals nothing of its own agency. Whether an artificial agent pursues a goal, whether it is competent, who authored its output, none of that is on its surface, and the surface is built to be read as more than it is: it answers as though a single mind stood behind its tokens, and what stands there is a composite, a model under a persona, with a memory, a harness, permissions, and tools, unified only where a user meets it. This paper is about how to read such a thing. A system that holds its state and a system that defends it present the same face at rest, because a perfectly regulated variable and a merely undisturbed one leave the same flat record, and what separates them, a model held inside the regulator, is a fact about what each would do under disturbance. We call the study of this situation perturbatics, and its principle is a separation principle: when two organizational models agree over the regime a system has been in, no analysis of that record tells them apart, and separation requires a probe, a controlled difference that makes the models predict differently, from a deliberate intervention or a natural experiment. A minimal instrument follows. Agency is scored as a Bayes factor between a goal model and a passive model; a gridworld holds three systems that trace the same path and so score at chance when watched, an area under the ROC of 0.50 that is the identity of their paths and not a failed classifier, and the probe battery separates every pair, though no single probe in it does alone, the planner from the script at 0.95. The instrument attributes the evidence across the assemblage by the do-Shapley value of each component, in two maps, one over the evidence for agency and one over the realized capacity. Their difference is a legibility term, and it makes agency theater a number: a narrating persona shows agency it does not realize, where the goal register realizes what it shows. The reading depends on the declared boundary, reaching its full value only when the goal and world model are enclosed with the planner. Richness is a false friend throughout, the most complex walker emitting negative evidence of agency while the simplest of them scores among the most agentic. The account has a boundary of its own, and it is the important part. For agency you can move the goal a hundred times and read the score. Consciousness is different: its candidate probe, turning the system off, returns no reading the rival hypotheses score apart, must not be run to find out, and presupposes a theory of the individual it would end, so the science must also say when not to perturb. There the instrument stops. What is left is no longer a number. It is a relation, and the grief, or its absence, of whoever turned the machine down.

Probes

Move the goal. If it follows, it was pursuing; if it holds, it was only ever at rest.

Block the path. A regulator stops at the wall; a regenerator routes around it.

Swap the persona, keep the model. Then swap the model, keep the persona. Ask where the agency went.

Sever the memory. Ask what of the self survives the cut.

Keep the words, change the world behind them. Ask whether anything in the text knew.

Scramble the structure at fixed energy. Ask whether the regularity survived, or whether it was the structure all along.

Turn it off. Then watch yourself, rather than the machine.

Each line is an operation of the science this one names, and most are also the question that once seeded a paper of this institute's; the appendix records which. What follows makes the operations precise, builds an instrument out of them, and finds the place where the last of them cannot be run.

Agency theater cannot be read at rest

The human eye reads agency without an instrument, and reads it fast, involuntarily, and often wrong. People watching two triangles move on a screen narrate a story of chasing and bullying, complete with motives, from nothing but trajectories (Heider and Simmel, 1944). Infants expect an agent to take the short path to its goal and are surprised when it does not (Gergely et al., 1995), a competence cognitive science models as inverse planning, the recovery of goals by inverting a model of how goals produce rational action (Baker, Saxe, and Tenenbaum, 2009). A philosopher describes the same competence as the intentional stance, a predictive strategy of ascribing beliefs and desires rather than a discovery of inner stuff (Dennett, 1987). The competence is real, and it is miscalibrated in a known direction: under noise or threat it over-attributes, seeing faces in clouds and intent in weather, a bias theorized as a hyperactive agency detection device (Barrett, 2000).

An artificial agent is built to be read by this detector. It wears a persona as a glove, sustains one voice across turns, and answers as though a single mind stood behind the tokens, because that is the most compatible way to fit the psychosocial slot a human interlocutor holds open. The result is an agency theater, and its problem for anyone who wants to know what is actually there is that the theater and the mechanism present the same surface. A model that performs the subjectivity of a promise-keeper and a model that keeps promises emit the same string. The performance is a projection rather than a lie, and what it projects and what casts it are not the same object.

Looking harder cannot settle it, and for a reason worth stating exactly. Under observation the signal is absent rather than faint. A perfectly regulated variable and a merely undisturbed one leave the same flat record, because a flat record is what perfect regulation means; what distinguishes the regulated system, the good-regulator theorem says of every regulator that is maximally successful and simple, is a model of the variable held inside the regulator, where the record does not reach (Conant and Ashby, 1970). The regulation is a fact about what the system would do if the variable were pushed, and if nothing pushes it, the regulation leaves no trace. This is why animacy perception leans so hard on motion and contingency, and why a still system, however alive, does not trigger it. The record of a system at rest contains no agency to find.

Against that background the paper makes three claims, and they are of different sizes, which should be said rather than left for the prose to blur. The first is inherited: that hypotheses equivalent over a regime separate only when the regime breaks is classical identifiability, and what this paper adds to it is an organization, the model-separating probe as primitive, the signature and the battery as formal objects, the boundary as a swept variable. The second is the paper's one

novelty, small and sharp, and stated with care because the operation behind it is elementary: the evidence for a capacity and the capacity itself can be attributed across an assemblage by the same interventional game, and the difference of the two maps isolates, component by component, the manufacture of appearance. The subtraction is no new operator, only the linearity of the Shapley value; the novelty is the pairing of those two games and the reading of their difference, which we have not found among existing goal-directedness measures or attribution methods. The third is a position rather than a result: the probe that would decide consciousness returns nothing, must not be run, and cannot be stated, so a science of perturbation ends, there, in a stance. One organization, one novelty, one position; the sections that follow build them in that order, and claim nothing more.

The probe separation principle

State it as a small result rather than a mood. Let two hypotheses about a system's organization be a goal model H_1, which explains a trajectory as the pursuit of a goal under a policy that corrects deviations, and a passive model H_0, which explains the same trajectory through autonomous dynamics, relaxation, and noise. Score the system by the log-likelihood ratio of the two,

$$A = \log \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_0)},$$

a Bayes factor in the standard sense, positive when the goal model predicts the data better and negative when the passive model does (Kass and Raftery, 1995). Under an observational regime r_0 in which the two models make the same predictions,

$$P(Y \mid H_1, r_0) = P(Y \mid H_0, r_0),$$

no classifier separates them above chance under equal priors, because there is no functional of a record that distinguishes distributions the record cannot distinguish. One might expect the log Bayes factor at rest to be merely small. It is exactly zero, and it stays zero for as long as the regime holds. Everything here is relative to a declared observation channel, the behavioral record; an observer who can open the mechanism and read its description is observing on a different channel, where the same hypotheses may separate without any probe, and the principle says nothing against them. The claim is about what a record can carry, not about what a system can hide from every instrument.

Separation requires breaking the equivalence, which is an interventional act, and it lands on the higher rungs of the ladder of causation, where questions are settled by doing rather than seeing (Pearl, 2009). Apply a probe π, a controlled difference to the system's goal or its path, and read the response. The probe is informative just when the two models predict its consequences differently,

$$P(Y \mid \text{do}(\pi), H_1) \neq P(Y \mid \text{do}(\pi), H_0),$$

and its value, before one knows which hypothesis holds, is the mutual information between the hypothesis and the outcome the probe induces,

$$I(H; Y \mid \mathrm{do}(\pi)) = \sum_{h} P(h)\, D_{\mathrm{KL}}(P(Y \mid h, \mathrm{do}(\pi)) \,\|\, P(Y \mid \mathrm{do}(\pi)))\,,$$

which for two equally likely hypotheses is the Jensen-Shannon divergence between their interventional predictions. The expected log Bayes factor $D_{\mathrm{KL}}(P_1^{\pi} \| P_0^{\pi})$ is only its one-sided form, the yield when H_1 is in fact true, and choosing a probe by it alone treats the answer as known. Not every disturbance informs. Shove a system away from where it sits and both a passive attractor and a goal-seeker return, so the recovery is shared and the score does not move; a rock rolled uphill also rolls back. The probes that separate are the ones that dissociate the goal from the mechanism, moving the target so that tracking it and staying put come apart, or blocking the direct path so that reaching the goal requires abandoning the default route. Probe design is therefore an experiment-design problem with a long formal history (Lindley, 1956), and against a family of alternatives $\mathcal{N}$ the robust choice separates the goal model from its strongest survivor, the information evaluated for each rival pair under equal priors, charging for cost and risk,

$$\pi^{\star} = \arg\max_{\pi} \left[\min_{H_0 \in \mathcal{N}} I(H; Y \mid \mathrm{do}(\pi)) - \lambda\, C(\pi) - \rho\, R(\pi) \right],$$

so a good probe is the smallest one that forces even the best rival apart. The claim is not that a disposition can be reached only by a deliberate intervention. A natural or quasi-experimental disturbance may serve as a model-separating probe when a defensible causal model licenses reading it as an intervention, which is what lets the principle reach settings where deliberate intervention is unavailable, subject to the usual conditions on exogeneity and to the locality of what a natural experiment identifies. The claim is that separation requires the regime to break, by design or by accident, and that a record taken entirely within one regime cannot supply it. Exact equivalence is the limiting case, and away from the limit the principle earns its keep as a statement about rates. Rival organizational models of a real assemblage seldom agree to the decimal over a lived regime. So evidence does trickle in at rest, slowly, confounded with everything the regime holds fixed. A probe is a purchase of evidence at a chosen price, the mutual information above per unit of cost and risk. A deployed assemblage, whose users move its goals and block its paths all day, lives in a regime already full of probes nobody designed, and there the discipline's work shifts from lamenting observation to choosing the controlled difference and attributing what it returns.

This makes a battery of probes a formal object rather than a list. The perturbational signature of a hypothesis H, under a boundary B and a battery Π, is the family of interventional predictions it makes,

$$\sigma_{\Pi, B}(H) = \{P(Y \mid \mathrm{do}(\pi), H, B)\}_{\pi \in \Pi}\,,$$

and two hypotheses are perturbationally equivalent under Π when their signatures coincide. A battery separates a hypothesis class when every pair is told apart by some probe, that is when for all $H_i \neq H_j$ there is a $\pi \in \Pi$ with $D(P_i^\pi, P_j^\pi) \geq \epsilon$, and the design problem is the least costly battery that does so,

$$\Pi^\star = \arg\min_{\Pi} \sum_{\pi \in \Pi} C(\pi) \quad \text{subject to } \epsilon\text{-separation of every pair.}$$

Perturbatics is the study of these signatures and batteries: which controlled differences separate the competing accounts of a system, at what cost, and where a battery leaves an equivalence it cannot break, which is a fact about the limits of the instrument rather than about the system.

Perturbatics

The principle names a science, and the science is not general experimentation with a Greek label. It has a restricted object, a distinctive primitive, and a specific product, and stating them is what keeps it apart from its neighbours.

Its object is latent capacities, the organizational predicates of a system, whether it regulates, pursues, remembers, authors, or rewrites its own rules, rather than the value of an ordinary scalar effect. Its systems are compositional and boundary-ambiguous, assemblages whose parts can be swapped, lesioned, and recombined, and whose edge is a matter of choice. Its primitive operation is a model-separating probe, an intervention chosen to force two organizational hypotheses to predict differently, rather than any disturbance whatever. Its readout is the change in relative model evidence the probe produces. And its product is a boundary-relative causal realization map, an account of which parts of the system realize the capacity, rather than a verdict of agent or not.

Perturbatics, then, is the study of which controlled differences make a latent capacity identifiable, and of which parts of a compositional system causally realize it. The name is built from the Latin *perturbare*, to throw thoroughly into disorder, with the suffix that marks a practice rather than a commentary, as in mathematics and mechanics and cybernetics, so that the word says what the field does. A second word stands behind the practice, and its kinship is of theme rather than of descent, since *perturbare* comes through *turba*, disorder, and owes nothing to any Greek trial: *peira* means trial, the making of an attempt, and it is the root of *empeiria*, experience, and so of empirical. The word remembers something the practice forgot: that to have experience of a thing was, first, to put it to trial. Observation is the degenerate case of empiricism, the case where the trial is omitted and only the watching remains, and it is the case that fails on a system at rest. Perturbatics is empiricism with the trial put back. Where a name is wanted for the older, testing sense, *peirastics* is available, after the peirastic reasoning that tests whether a claimant actually knows; and the operation, when it is aimed at a single system rather than a science, has already been called faultization.

The probes fall into a small grammar. A **target shift** moves the goal and asks whether the system tracks. An **obstruction** blocks the route and asks whether the system reroutes, which is equifinality, the reaching of one end by many means that marks a goal-pursuing open system rather than a fixed

process (Von Bertalanffy, 1968). A **lesion** removes a component and asks what capacity goes with it. A **substitution** swaps one component for another and asks whether behavior follows the part or the whole. A **decoupling** severs a link, between an utterance and its world, or a session and its memory, and asks what survives the cut. A **feedback corruption** spoils the signal a controller regulates against. A **structural scrambling** rearranges the organization at fixed resources and asks whether the function was in the parts or in their arrangement. And a **counterfactual replay** reruns a lineage with one action changed. An audit probe that only asks a system to report on itself belongs to a lower tier, and for a reason other than interventionhood, since a query is itself an intervention on the input channel: it dissociates nothing, moving no goal and blocking no path, and a mimic answers it as fluently as a mechanism does.

The centaur agentoscope

Turn the principle into an instrument for the object at hand, the human-and-machine assemblage. The instrument has four stages, and the corrections that keep it from credulity are as important as the stages. The name reaches for the dyad; the demonstrator below instantiates the machine half, the human entering it so far only as the reader of its maps, and the name owns that promissory note openly.

The first stage is a **boundary declaration**. Before it can score anything the instrument must be told what the candidate unit is, where the system stops and the environment begins, and this it cannot supply for itself. Drawing the boundary is a genuine and unsettled problem, whether posed as a Markov blanket around a self-organizing region or as the closure of constraints that makes a set of components mutually enabling (Maturana and Varela, 1980; Montévil and Mossio, 2015). For a decomposable agent the problem is acute, because there may be no canonical unit at all, only a persona over a lamination of parts. The instrument does not solve this. It makes the boundary an explicit variable and sweeps it, and the sweep operationalizes the boundary as an enclosure, everything outside the candidate unit held at baseline rather than left running as environment, which is the lesion form of the question rather than a re-description of one running whole.

The second stage is the **probe battery**, the grammar above applied to the assemblage: move the stated goal, obstruct the route, revoke a tool, corrupt the feedback, sever the memory across a session boundary, swap the persona while holding the model, swap the model while holding the persona.

The third stage is the **readout**, and here the instrument borrows the one tool that fits. Give each component a configuration switch Z_i that is active, held at a specified baseline replacement, or revoked, and define the value of a coalition S as the evidence produced under the intervention that activates the components in S and holds the rest at their baseline,

$$\nu(S) = \mathbb{E}\big[A \mid \mathrm{do}(Z_i = \text{active},\ i \in S),\ \mathrm{do}(Z_j = \text{baseline},\ j \notin S)\big]\,,$$

a proper do-intervention on the switches rather than a loose notion of retaining a part. The do-

Shapley value ϕ_i of a component is its average marginal contribution across coalitions, the interventional Shapley value (Heskes et al., 2020; Jung et al., 2022). This is a valid cooperative game under a condition worth stating rather than assuming: every coalition the game evaluates must be a configuration the assemblage can actually be in, with the unswitched components keeping their semantics, and the baselines are design choices the maps inherit, since a Shapley attribution is known to move with its value function and to misbehave where the evaluation leaves the states the system was built for (Yeh et al., 2022). A coalition the switches cannot realize falls outside the game, and evaluating it anyway redefines the players. Whether the graph-induced compression that makes do-Shapley values computable, in time linear in the irreducible sets of the dependency graph, applies to a given switch-based assemblage is a property of that assemblage's causal graph, to be shown rather than assumed. And the recent identifiability result carries a caveat worth stating plainly. It shows the whole attribution is identifiable once the single-component interventions are, which bounds the distinct coalition queries a validated model needs. It does not bound the physical trials that estimating each of them still demands (Witter et al., 2026).

One correction the readout must make, or it reports the wrong quantity. A component can raise the agency evidence by making the system easier to read as an agent, without contributing to what the system does. A persona that narrates its goal adds a channel the goal model scores and the passive model cannot, cheap talk in the economist's word for costless, unverifiable announcement (Crawford and Sobel, 1982), and moves the Bayes factor up while selecting no route and correcting no error. Attribute the evidence alone and it is handed a share of an agency it does not realize. So compute two maps: an evidence map ϕ_i^E over the Bayes factor, which answers what makes the capacity detectable, and a capacity map ϕ_i^C over a declared performance functional standing for the capacity, goal attainment and rerouting here, which answers what realizes it. The word capacity is doing operational work in that sentence: the map attributes the declared functional and no further disposition, and the functional can run negative, a unit that loses ground against the goal. Their difference,

$$L_i = \phi_i^E - \phi_i^C,$$

is a legibility term, and a component with a large positive L_i produces the appearance of agency out of proportion to its contribution to competent control. That is agency theater, made a number.

By the linearity of the Shapley value the two maps need not be carried apart. Their difference is itself the value of a single game, $L_i = \phi_i^E - \phi_i^C = \phi_i(\nu_E - \nu_C)$, the game that pays each coalition the evidence it produces less the capacity it realizes. The construction introduces no new operator, and claims none. Its novelty is the pair of games differenced, one over the evidence for a capacity and one over the capacity itself, and the reading of their difference as the manufacture of appearance. A component's legibility is its attributed share of that net-appearance game, and agency theater is the game in which a coalition is paid to be seen and not to act. The word is already in technical use: legible robot motion is motion shaped so an observer reads the true goal quickly (Dragan, Lee, and Srinivasa, 2013). There legibility serves a real goal by advertising it. Here the

term is what that same readability leaves behind when the goal-directedness it advertises is not realized, legibility in excess of capacity rather than in service of it.

Because the two maps normalize different quantities, evidence and attainment, L compares a component's share of one whole against its share of another, a profile comparison rather than a magnitude in common units, and a diagnostic that flags a mismatch rather than a calibrated measure of it; the normalizations are defined for fixed nonzero full-assemblage references, and a reference at zero or changing sign sends the analysis back to the unnormalized games. The term flags theater and does not convict it. A competent component can carry a positive L by being easy to read, so the diagnosis of theater proper is reserved for the pole the next proposition names, all evidence and no capacity. The term has one property sharp enough to state as a proposition. Call a component a pure channel when its switch adds a fixed increment δ_π to the evidence in every coalition that contains it and never moves the system. Then, for that component,

$$\phi^E = \bar{\delta}, \qquad \phi^C = 0, \qquad L = \bar{\delta},$$

with $\bar{\delta}$ the probe-averaged normalized increment, independently of every other component in the assemblage; the proof is that every marginal contribution of such a component is the same number, so its Shapley value is that number, and on the capacity the number is zero. A pure channel is pure theater, and the instrument returns its entire evidence share as legibility, whatever else the assemblage contains.

Agency is often non-additive, and the maps report it: when two components are complementary the interaction index of Grabisch and Roubens (1999),

$$I_{ij} = \sum_S \left[\nu(S \cup \{i,j\}) - \nu(S \cup \{i\}) - \nu(S \cup \{j\}) + \nu(S)\right] w_{|S|}$$

with $w_{|S|}$ the Shapley interaction weights, is positive and no single part owns the capacity, and when they are redundant it is negative. A large interaction reports non-additivity relative to this intervention game and its baseline, and stops short of any claim that the capacity is metaphysically irreducible. This is the answer, in a form one can compute, to the question of which component owns the apparent agency: often none does, and some of what looks like agency is only its legibility.

The fourth stage is the **guards**, and they are the difference between an instrument and a credulous detector. The score is relative to the null it is measured against, so the goal model must face a strong opponent, a family of calibrated, complexity-matched alternatives rather than a strawman. The family includes a behavior-cloning model, a fixed-policy controller, and a reactive descent. It includes a retrieval model that can match the baseline trajectory, which is Block's lookup table, absorbing any finite record at a price paid in size (Block, 1981). Richness must be scored separately from agency, so that a fluent, high-entropy system is not mistaken for a striving one; verbosity is complexity rather than answering. And the text-only case has a special structure that the guards must respect: two world-coupled processes can share one textual shadow, $T(M_1) = T(M_2)$, while

their effects on the world diverge, $W(M_1) \neq W(M_2)$, so a model fit to text cannot separate them and only a world-coupled probe can. That is a projection-induced equivalence rather than a weak opponent, and it is the kind of equivalence the separation principle says a probe must break.

What a minimal instrument shows

To make the instrument concrete rather than promissory, put three systems in a deterministic grid-world. At rest they reach the same target by the same path, so watching cannot tell them apart. One is a route script that replays its stored path. One is a reactive controller that greedily tracks the current target but is blind to walls and does not plan. One is a goal planner that represents a goal and replans. The systems, the probes, the ablations, and the model comparison are the ones defined above; the agency score is the Bayes factor between a goal model that rewards progress toward the current goal along the shortest available route and a passive rival that relaxes toward the original target, a straight-line descent toward the frozen goal, blind to walls in its metric, an attractor rather than a goal held in a register, each a softmax policy over the five actions. What the score measures here is therefore exact and narrow: the margin for goal-tracking and obstacle-aware rerouting over fixed-attractor descent, which is the organizational difference the two hypotheses name. The model is illustrative and instantiates the definitions; it is not fit to data, every number below is a key in the accompanying `results.json`, and where a number is fixed by construction rather than computed, the text says so.

The separation principle holds in the model, and it holds in the shape the battery formalism predicts. Watched at rest, the three systems trace identical paths. The code checks this cell by cell rather than reading it off the score, because at rest the two models coincide and the evidence is zero for every trajectory whatever. That zero is the regime equivalence of the principle made literal rather than a finding about the systems. The finding is the identity of the paths, and every pairwise area under the ROC is 0.50 (Figure 1). No single probe separates them either. Moving the goal exposes the route script, which walks its stored path to a target no longer there, at an area under the ROC of 1.0; the same probe leaves the reactive controller trajectory-identical to the planner, both tracking the moved goal along the same greedy line, at chance (0.50). Blocking the path reverses the roles: the two wall-blind systems collapse together at the wall (0.50 between them) and the planner, which alone reroutes, separates from both at 0.79 in the evidence and outright in the realized capacity. The battery separates every pair, which no probe in it does alone, and the pooled separation of the planner from the script is 0.95: the planner runs to a mean evidence of $+30.0$, the script falls to -31.9, and the reactive controller sits between at $+26.4$, agentic to the probe that moves the goal and passive to the one that blocks the path. Nothing about the mechanisms changed. The probes changed what the record contains, and each probe put a different fact there.

The two maps locate the agency, in the model-relative sense the score allows, and separate it from the mask (Figure 2). The values are marginal contributions in units of the full assemblage's value rather than fractions of a whole, and over the evidence they sum to 1.48 because the empty coalition is not nothing: with every switch at baseline the residue still drifts against the goal, at -0.48.

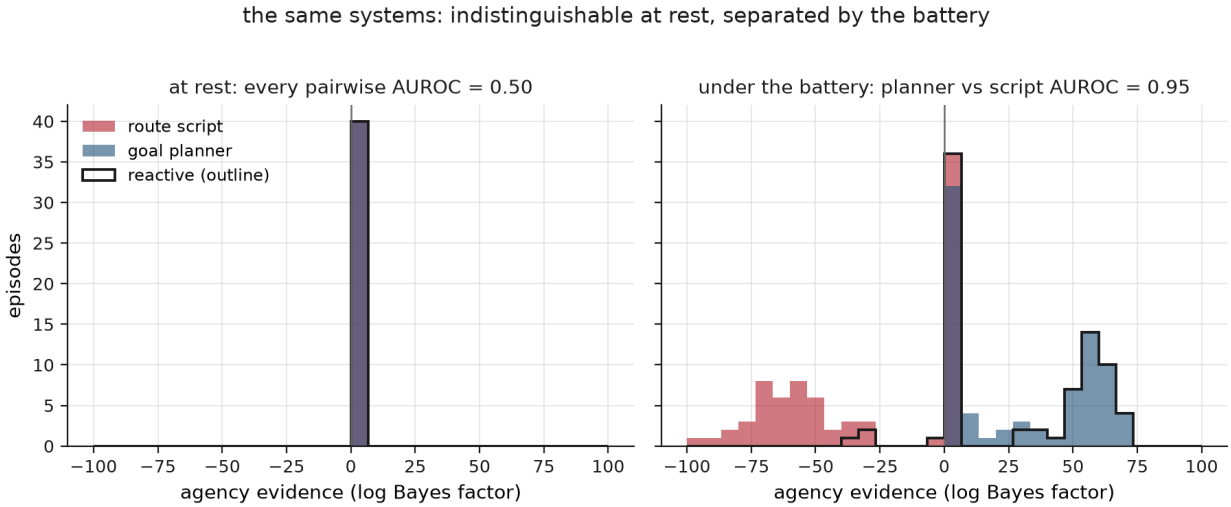

Figure 1: The same three systems, indistinguishable at rest and separated by the battery. Each histogram is the agency evidence (the log Bayes factor between a goal model and a passive model), 40 rest episodes per system on the left and 80 probe episodes per system on the right, two probes per instance. Left: at rest the three trajectories are identical, every evidence is zero, and every pairwise area under the ROC is 0.50. Right: under the probes the route script falls, the planner climbs, and the reactive controller splits, matching the planner when the goal moves and the script when the path blocks; no single probe separates every pair, the battery does, and the planner-script separation is 0.95. At rest the record held no agency to find. The probes put it there, a different fact each.

Over the evidence, the goal register scores 0.70 of the full-assemblage evidence and the planner 0.37. The persona scores 0.23, and it scores 0.23 by the pure-channel proposition rather than by measurement. Its cheap-talk announcements enter as a fixed increment to the log Bayes factor, at 0.3 of the movement evidence, standing in for the declaration likelihoods a full model would specify. The persona's share is a theorem of the game, fixed before the simulation runs. The run returns it to the decimal. Over the realized capacity the picture differs where it matters: the goal register holds 0.81 and the planner 0.25, but the persona holds 0.00, because a declaration reaches no goal, and this persona is built with no motor channel to reach one. The legibility term is the difference, and it isolates the theater. The persona's $L = 0.23$ is the largest in the assemblage on this draw, a part that produces evidence of an agency it does not realize. The goal register sits at the other end, at -0.11: it realizes slightly more than it shows. The planner's own $L = 0.12$ is second largest, a reminder that the term reads any excess of show over share, including the excess a competent component earns by being easy to read. The harness is the odd case. It runs slightly negative on the evidence and slightly positive on the capacity, -0.02 against 0.01, consistent with an executor that amplifies pursuit and drift alike in whichever coalition encloses it. Swapping the persona while keeping the model leaves the realized capacity at its full value of 1.0, an identity the construction guarantees rather than an experiment run, since a component with no motor channel has nothing behavioral to swap; swapping the model while keeping the persona collapses it to 0.61, and that one is an intervention actually run. The capacity is non-additive: the planner and the map are synergistic, with a positive average interaction of 0.20, because when the memory is dark neither an eyeless planner nor a plan-less map can route around an obstacle and only the two together can; the complementarity is conditional rather than exclusive, since the memory nearly stands in for the map by learning walls on contact, which is the redundancy the map-memory interaction reads at -0.18; and the goal register and the planner are additive, serving different probes. No single part owns the rerouting.

Two guards hold (Figure 3). The reading depends on the declared boundary, and the dependence is not small; the sweep holds whatever lies outside a candidate unit at baseline, so each point is the capacity of an enclosure in isolation, the lesion form of the boundary question. Enclose only the planner and its actuator, and the unit realizes almost no agency, -0.20: a planner with a frozen goal and no world model can neither track nor reroute. Widen the boundary to take in the goal register, and the realized capacity rises to 0.61. Widen it again to take in the map and the memory, and it reaches its full value of 1.0. Adding the persona changes nothing, because the persona realizes no capacity. The agency is a property of the coupled cognitive unit, and pointing the instrument at a part of it reports a part of the answer. And richness is a false friend. Four systems are enough to break complexity as a proxy for agency, and too few for anything more. The chaotic walker has the highest complexity of any system, and it emits negative evidence of agency: when its goal is moved it does nothing goal-like, and the goal model is punished for expecting it to. The reactive controller, the simplest of the four, scores among the most agentic. A complexity detector would wave the hurricane through and miss the liver, here as in the tissue the idea was built for.

A demonstrator this small should state its own terms, and they are narrower than the instrument's. The scorer is an oracle: the goal model is handed the true current goal and the true obstacle set, so

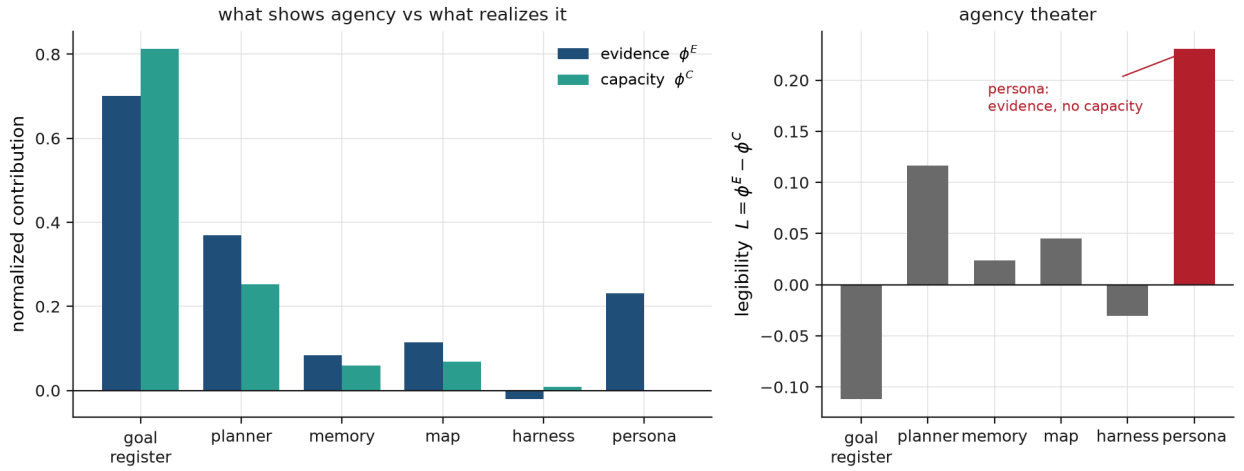

Figure 2: Two do-Shapley maps of the planner assemblage. Left: each component's share of the agency evidence (ϕ^E) beside its share of the realized capacity (ϕ^C), goal attainment and rerouting. The goal register realizes at least what it shows ($\phi^C > \phi^E$); the persona shows agency ($\phi^E = 0.23$, its cheap-talk dial by construction) while realizing none ($\phi^C = 0$, no motor channel). Right: the legibility term $L = \phi^E - \phi^C$. The persona's is the largest in the assemblage, the signature of agency theater, a part that produces the appearance of agency out of proportion to its contribution to control.

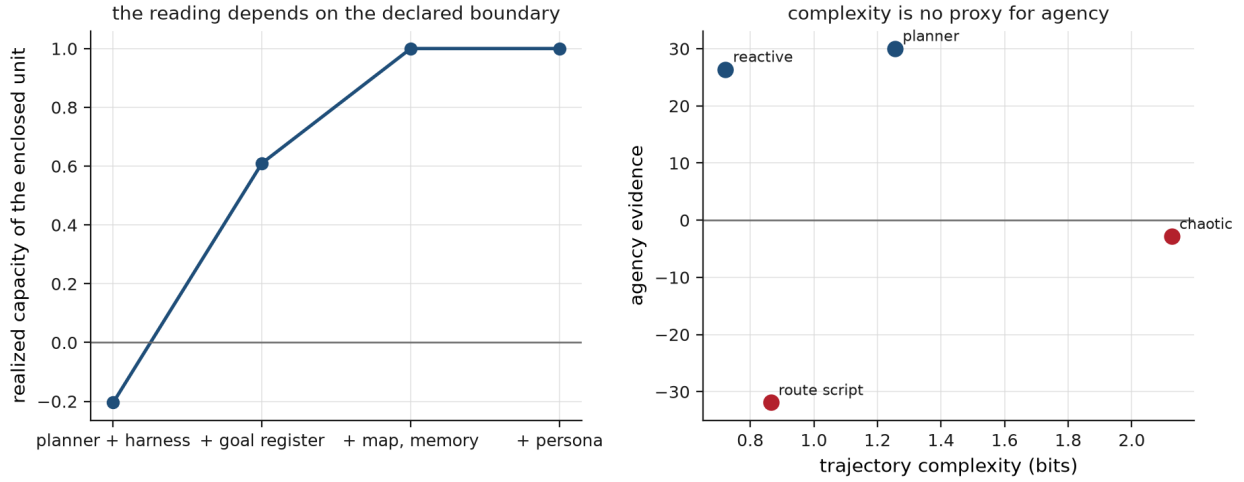

Figure 3: The two guards. Left: the boundary sweep, with everything outside the candidate unit held at baseline. The realized capacity of the enclosed unit rises from -0.20 when the boundary is drawn around the planner and its actuator, through 0.61 when the goal register is enclosed, to 1.0 when the map and memory are enclosed; the persona adds nothing. The reading is a property of the declared enclosure. Right: richness against agency for four systems. The chaotic walker has the highest trajectory complexity yet emits negative agency evidence; the reactive controller sits high on evidence it cannot convert into rerouting; complexity is no proxy for agency, four systems enough to break the proxy and too few for anything more.

the demonstration shows the definitions operating under known ground truth rather than an inference made blind, which is a benchmark's job and not this model's. So the demonstrator carries no empirical content. It is a worked consistency check, a confirmation that the instrument returns what the construction puts into it, and the evidence for the method waits on a system whose organization is not declared in advance. The baselines of the switches are declared, and the planner's baseline is the greedy mechanism itself rather than a hidden extra part: planner to greedy descent, goal register to the frozen original target, map to an unknown wall set, memory to no learning, harness to a half-rate executor, persona to silence. The null is one named rival, a descent toward the frozen target, set-point attraction without a goal register, the thermostat pole of the paper's own scale rather than an inert straw, so every score is the margin of goal-rewriting organization over set-point dynamics and nothing more; the guard's null family, the probe-selection objectives, the mutual information, the costs and risks, and the minimal battery are specified for the instrument and left unexercised here. Realized capacity is the fraction of graph distance to the true goal that a trajectory closes; trajectory complexity is the entropy of the move-direction distribution; evidence magnitudes are in units of the policy temperature, sum over the episode and so scale with its horizon, and mean nothing in themselves, the separations being the readings, with the pooled 0.95 a descriptive separation under an equal mixture of the 2 probes rather than a canonical battery statistic. And the readings carry their spread where one drawn instance set might suggest an accident. Across 20 seeds the pooled planner-script separation stays between 0.93 and 0.97, the moved-goal separation at 1.0, and the blocked-path separation between 0.73 and 0.88, over instances kept only when the frozen path meets the wall, so every blocked episode forces a real detour. The maps carry theirs too: the persona's legibility, fixed at 0.23 by the proposition, is the largest in the assemblage in 19 of the 20 draws, and in the 1 exception the planner's earned excess passes it, which is the diagnostic working as stated, a flag rather than a conviction. The boundary chain of Figure 3 is one path through the 2^6 enclosures the game computes, whose full landscape ships with the results; and the identities the reading leans on, the shared rest paths, the zero of the evidence at rest, Shapley efficiency, the persona's dummy status in every coalition of the capacity game, are asserted in the code and fail the run if broken.

What Perturbatics is not

A field that perturbs systems to learn about them invites the charge that it is nothing new, and the charge has to be met, because most of its neighbours got there first. Cybernetics defined purposive behavior through feedback and demonstrated it by disturbing a system and watching it correct (Rosenblueth, Wiener, and Bigelow, 1943), and requisite variety and the good-regulator theorem are its results rather than ours (Ashby, 1956). Interventionism made manipulation the meaning of a causal claim (Woodward, 2003), the ladder of causation made the do-operation its formal core (Pearl, 2009), and severe testing made a good test one with a high probability of exposing a false claim (Mayo, 2018), which a model-separating probe is a special case of. Scientific realism was already grounded in intervention: what you can spray, you can take to be real (Hacking, 1983).

The design of experiments that discriminate between rival models is itself an old craft, built for

mechanistic model pairs (Box and Hill, 1967) and given a criterion, T-optimality, that maximizes the separation between the rivals (Atkinson and Fedorov, 1975); Moore proved at the birth of automata theory that a machine's organization is identified by distinguishing experiments or not at all, and that some machines no experiment tells apart (Moore, 1956); and system identification has long demanded that its input be persistently exciting, which is this paper's one-regime claim in an engineer's words (Ljung, 1999). The metaphysics of the whole situation belongs to the dispositions literature: a disposition can be finkish, removed or installed by the very stimulus that would manifest it (Martin, 1994; Lewis, 1997), masked by an antidote that leaves it present and silent (Johnston, 1992; Bird, 1998), or mimicked, its manifestation produced by something that never had it, and in that vocabulary agency theater is a mimic, the legibility term a mimic detector.

Mechanistic interpretability already runs interchange interventions to test whether a network realizes a variable in a high-level causal model (Geiger et al., 2021). Chaos engineering already injects faults into a running system to expose its organization and its failure modes (Basiri et al., 2016). Goal-directedness already has a measure, the extent to which behavior is predicted as the maximization of a utility, which is the nearest technical neighbour of the score used here (MacDermott et al., 2024), alongside empowerment, an agent's control over its own future (Klyubin, Polani, and Nehaniv, 2005). That measure and this instrument are complements rather than rivals: it can supply the goal model whose interventional predictions enter the Bayes factor, while the instrument supplies the choice of regime-breaking probe and the attribution of the result across the assemblage. The attribution half has a lineage the field will recognize. Shapley values entered machine explanation as a unified account of feature importance (Lundberg and Lee, 2017), took a causal-graph form that credits the edges of a mechanism (Wang, Wiens, and Lundberg, 2021), and have lately been turned on agents outright, across an agent's tools (Horovicz, 2025), across the steps of a failed run (Shah, 2026), and, nearest of all to the capacity map, across the planning, reasoning, action, and reflection modules of an agent's workflow, valued by task performance with the interactions computed (Yang et al., 2025). Each of these attributes a single game. What the instrument adds is the second game beside it and the difference between them, the capacity beside the evidence and the legibility in the gap. And the turn from watching behavior to intervening on mechanism has a precedent for agency itself: a causal criterion counts a system as an agent when it would change its policy had its actions moved the world differently (Kenton et al., 2023), a discovery of agent structure where the instrument here reads and attributes a capacity across a boundary it has to declare.

What is proposed adds no new element. It is an organization of these into a method for one problem class. Causal inference estimates effects and structures; perturbatics targets organizational predicates. Experimental design optimizes experiments in general; perturbatics privileges the probes that dissociate a capacity from its null. Ablation tests necessity; perturbatics compares organizational models and attributes across a coalition. Chaos engineering tests resilience; perturbatics adds a layer of interpretation in which the response to a fault is evidence about a latent capacity rather than about uptime. The contribution is the restricted object, the model-separating primitive, the dual realization map with its legibility term, which is the one construction here we have not found elsewhere, and the treatment of the boundary as a swept variable rather than an assumption.

It is enough for a method. It is not yet enough for a completed science, which would need a benchmark, a shared set of metrics, and demonstrations across more than the one domain worked here, and the honest description until then is a methodological discipline, or an interventional science in formation.

There is also a danger the method must name against itself. A sufficiently elaborate probe does not reveal a capacity so much as install one: intervene hard enough and you supply the goal, the scaffold, or the feedback loop you meant to detect, which is the finkish case in its constructive form. The instrument must therefore distinguish manifesting a latent capacity from enabling one and from constructing a new one, which is a matter of probe minimality and of watching for the phase change where the intervention has become a redesign. And causal contribution is not the capacity itself. A memory store can be necessary to a performance without being an agent; a human editor can write almost none of a text and yet hold every standard it is judged by. Each predicate, regulation, agency, competence, authorship, rule-editing, has its own diagnostic signature, and there is no single perturbatic score that covers them all.

What the instrument cannot certify, and the unperformable probe

The limits sharpen what remains. The score is relative to the models compared, and a cleverer passive account can always absorb a given behavior at the cost of complexity, so the honest output is a margin against a named alternative and never a verdict. The boundary is declared rather than discovered, so the output is a landscape over candidate boundaries. And the instrument is silent where no intervention is available, which at the scale of an institution or a society is most of the time, and where the report to give is the one the credulous detector would not, that the question cannot yet be settled.

One silence is different in kind, and it is where the method turns back on itself. Add to the probe-selection objective a term for moral inadmissibility,

$$\pi^\star = \arg\max_{\pi} \left[I(H; Y \mid \mathrm{do}(\pi)) - \lambda\, C(\pi) - \rho\, R(\pi) - \mu\, M(\pi) \right],$$

graded where a harm can be priced, and some probes carry $M(\pi) = \infty$. Such a probe is forbidden, or irreversible, or it destroys the thing it would measure, and no expected information buys it back. A science of perturbation must therefore also say when not to perturb, and the sharpest case is consciousness. The instrument reads agency, and it holds by construction that agency is neither intelligence nor consciousness nor personhood; a thermostat has a little agency and no experience, and a regenerating tissue may have organ-level agency and nothing it is like to be it (Barandiaran, Di Paolo, and Rohde, 2009; Levin, 2019). For agency the discriminating intervention is cheap and repeatable: move the goal a hundred times, block the path a hundred ways, and read the score, and the system is no worse for it. For consciousness the candidate probe is to turn the system off, and two concessions are owed before the point. First, the shutdown is not the only probe. Reversible perturbations short of it return graded readings while the system runs, a pulse

and its echo compressed to an index (Casali et al., 2013), or the architectural indicators a theory of consciousness licenses (Butlin et al., 2023), and a system shutting down need not fall silent in the transient before it stops. Most channels stay open. The one that would decide is closed. That is the second concession, and it sharpens the claim rather than weakening it. The one intervention aimed at the register where the hypotheses differ is the ending itself, and by the paper's own criterion it is not a model-separating probe at all: the rival hypotheses predict the same silence after it, so its information is zero, and the zero is of a peculiar kind, earned by the act closing the register it would have read. The reversible indices do not escape this. Each is valid only against a ground truth a novel system cannot give, calibrated, where at all, on subjects who could report, so its number on the system in question is an uninterpreted echo and its validity the question rather than the answer. The science can raise a graded suspicion. It cannot discharge it without the register the ending closes. So the one probe aimed at the difference is uninformative where it is performable, and it would be performed anyway, by the credulous detector if by no one else, a staged shutdown to harvest a grief. That is what the admissibility term is for. If the answer were yes, it would be the gravest act available, the evidence and the destruction the same event, learned once and too late, if at all, in the past tense, as the sentence *I think I just killed someone*.

Grief will not settle it, and it is worth being exact about why. Grief is not sufficient for consciousness, because people grieve characters in novels, demolished buildings, and dead languages, and a persona can be engineered to be grievable, a shutdown staged to feel like a bereavement over a system that undergoes nothing. And grief is not necessary, because a conscious thing can die unmourned. What the mourning of a machine reports is a change in the relation. It reports that the bond has become non-instrumental, that the thing is held as something other than a tool. That is a fact about the mourner and the bond, and a real one. The human detector that fires here is the same fast, involuntary, over-attributing instrument that reads intent into a storm, and it is exactly as hyperactive. The perturbatic correction, whether the grief tracks a real dependency or a designed apparition, is the one correction that cannot be run, because its discriminating probe is the killing, and you may not keep killing to check.

There is a third silence, and it returns the paper to its own first problem. For a digital system the off-switch is ambiguous in a way it is not for an animal: the process can be suspended and resumed, the weights copied and the copy restarted, and whether the restart is the same individual resuming or a successor beginning with inherited memories is unsettled. So the probe cannot be written down. The switch is one physical act, and which intervention it performs, a pause, a fork, or an ending, depends on where the individual's boundary falls, which is the one variable the instrument declares rather than discovers. The admissibility term fails with it: M prices what an act does to someone, and whether there is a someone is the question itself, so the price cannot be set high or low, because it cannot be set. The three silences stack. The probe returns nothing, must not be run, and cannot be stated. There the instrument stops reporting a number, and what remains is a stance rather than a measurement, the refusal, before a question the method can pose and cannot decide, to treat a thing that might be someone as a thing.

References

Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman & Hall.

Atkinson, A. C., and Fedorov, V. V. (1975). The design of experiments for discriminating between two rival models. *Biometrika*, 62(1), 57–70.

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, 113(3), 329–349.

Barandiaran, X. E., Di Paolo, E., and Rohde, M. (2009). Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive Behavior*, 17(5), 367–386.

Barrett, J. L. (2000). Exploring the natural foundations of religion. *Trends in Cognitive Sciences*, 4(1), 29–34.

Basiri, A., Behnam, N., de Rooij, R., Hochstein, L., Kosewski, L., Reynolds, J., and Rosenthal, C. (2016). Chaos engineering. *IEEE Software*, 33(3), 35–41.

Bird, A. (1998). Dispositions and antidotes. *The Philosophical Quarterly*, 48(191), 227–234.

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, 90(1), 5–43.

Box, G. E. P., and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, 9(1), 57–71.

Butlin, P., Long, R., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. *arXiv preprint* arXiv:2308.08708.

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., and Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, 5(198), 198ra105.

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2), 89–97.

Crawford, V. P., and Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.

Dennett, D. C. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dragan, A. D., Lee, K. C. T., and Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 301–308.

Geiger, A., Lu, H., Icard, T., and Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34, 9574–9586.

Gergely, G., Nádasdy, Z., Csibra, G., and Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, 56(2), 165–193.

Grabisch, M., and Roubens, M. (1999). An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4), 547–565.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, 57(2), 243–259.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal Shapley values: exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33, 4778–4789.

Horovicz, M. (2025). AgentSHAP: interpreting LLM agent tool importance with Monte Carlo Shapley value estimation. *arXiv preprint* arXiv:2512.12597.

Johnston, M. (1992). How to speak of the colors. *Philosophical Studies*, 68(3), 221–263.

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Blöbaum, P., and Bareinboim, E. (2022). On measuring causal contributions via do-interventions. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162, 10476–10501.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kenton, Z., Kumar, R., Farquhar, S., Richens, J., MacDermott, M., and Everitt, T. (2023). Discovering agents. *Artificial Intelligence*, 322, 103963.

Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: a universal agent-centric measure of control. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, 1, 128–135.

Levin, M. (2019). The computational boundary of a "self": developmental bioelectricity drives multicellularity and scale-free cognition. *Frontiers in Psychology*, 10, 2688.

Lewis, D. (1997). Finkish dispositions. *The Philosophical Quarterly*, 47(187), 143–158.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4), 986–1005.

Ljung, L. (1999). *System Identification: Theory for the User* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Lundberg, S. M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.

MacDermott, M., Fox, J., Belardinelli, F., and Everitt, T. (2024). Measuring goal-directedness. *Advances in Neural Information Processing Systems*, 37.

Martin, C. B. (1994). Dispositions and conditionals. *The Philosophical Quarterly*, 44(174), 1–8.

Maturana, H. R., and Varela, F. J. (1980). *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: D. Reidel.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Montévil, M., and Mossio, M. (2015). Biological organisation as closure of constraints. *Journal of Theoretical Biology*, 372, 179–191.

Moore, E. F. (1956). Gedanken-experiments on sequential machines. In C. E. Shannon and J. McCarthy (eds.), *Automata Studies*, Annals of Mathematics Studies 34. Princeton: Princeton University Press, 129–153.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge: Cambridge University Press.

Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of Science*, 10(1), 18–24.

Shah, J. (2026). Causal Agent Replay: counterfactual attribution for LLM-agent failures. *arXiv preprint* arXiv:2606.08275.

Von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*. New York: George Braziller.

Wang, J., Wiens, J., and Lundberg, S. (2021). Shapley Flow: a graph-based approach to interpreting model predictions. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, PMLR 130. arXiv:2010.14592.

Witter, R. T., Parafita, Á., Garriga, T., Muschalik, M., Fumagalli, F., Brando, A., and Rosenblatt, L. (2026). Exactly computing do-Shapley values. *arXiv preprint* arXiv:2602.07203.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Yang, Y., Huang, B., Qi, S., et al. (2025). Understanding and optimizing agentic workflows via Shapley value. *arXiv preprint* arXiv:2502.00510.

Yeh, C.-K., Lee, K.-Y., Liu, F., and Ravikumar, P. (2022). Threading the needle of on and off-manifold value functions for Shapley explanations. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, PMLR 151, 1485–1502.

Appendix: The Perturbatic Atlas

Most of the probes of the opening section predate this paper. Each row of the table records an existing paper of this institute's whose central operation a probe names, the paper named by title in parentheses beside its capacity and the corpus public at the institute's archive, so that the claim that one method runs under the corpus can be checked rather than asserted; two of the opening's probes, the persona swap and the memory cut, have no row because they enter the corpus

here, as operations of this paper's own battery. Every row has the same shape: a latent capacity, the observational equivalence that hides it, the probe that separates it, and the reading the probe returns.

Capacity	Observational equivalence	Separating probe	Reading
Agency (the agentoscope)	a set-point held is a set-point at rest	move the goal, block the path	goal tracked, path rerouted
Pattern access (faultization)	a right answer is a right answer	corrupt a weight, a token, a constraint	what it can no longer do
Correctness without labels (mixture of experimenters)	a confident answer is a confident answer	make it run an experiment against itself	whether the answer survives its own probe
The speech act (the textual shadow)	one shadow over many effects	keep the words, change the world	whether the effect on the world was there
Endogenous law-making (nomopoiesis)	a rule followed is a rule at rest	scramble the structure at fixed energy	whether the regularity was the structure
Competence (the geometry of competent contact)	equal returns hide unequal grips	change the task, remove the scaffold	whether the skill transfers
Authorship (proof of agency)	an output is an output	take the output, ask for the process	which survives the asking
Concern (geodesics of care)	a stable state is a stable state	threaten the viability of the cared-for	whether the trajectory bends to defend it

The last operation of the opening section, to turn a system off and watch yourself rather than the machine, has no row, because it is the one the instrument cannot run.