The Agentoscope:
Why Agency Must Be Perturbed to Be Seen

PIATRA . INSTITUTE

July 2026

Abstract

Humans detect agency without instruments. A person reads pursuit in a tiger, intention in a baby, and effort in a struggling animal, and reads it fast, involuntarily, and often wrongly, in a storm or a shadow or a moving triangle on a screen. The proposal here takes that human competence as a prototype the way the eye was a prototype for the lens: the optics of agency-detection are already implemented in us, and the question is what the instrument is doing and whether it can be extracted, corrected, and enlarged into an agentoscope that finds agency where unaided perception cannot, most promisingly not outward in weather but downward in living tissue. This paper argues that the instrument is a model comparison. A system is agentic to the degree that its behavior is better predicted by a model with a goal, a boundary, and a policy that corrects deviations than by a passive model of forces and noise, and the agency score is the log-ratio of the two. The paper's central and somewhat unwelcome result is that this score is unreadable from observation alone. When a system's goal is not moving, the goal-model's predictions coincide exactly with those of a suitable passive dynamical model, so the two cannot be told apart by watching, and a simulated detector separates three hundred goal-seekers from three hundred passive attractors at an area under the ROC of 0.49, which is chance, with the agency score never once exceeding zero. Agency is not a property on the observational rung of Pearl's ladder. It appears only under intervention. When the goal is moved and the system tracks it, or the path is blocked and the system reroutes, the goal-model predicts what the passive model cannot, the score climbs by roughly 101 nats per informative intervention, and the same detector now separates the two populations perfectly. Not every perturbation informs: merely displacing a system, which a passive attractor and a goal-seeker both recover from, leaves the score at zero, so the discriminating interventions are exactly those that pit the goal against the mechanism. Two further results guard the instrument against the mistakes the human prototype makes. Dynamical richness is a false friend: across a battery of systems the agency score runs against complexity rather than with it, a chaotic flow scoring the lowest of all despite the highest complexity, which is why a hurricane is a false positive and a liver is not. And agency is graded rather than binary, ordered by the instrument into a ladder on which passive systems score below zero and goal-directed ones above, with equifinality under obstruction, reaching the same goal by a different path, separating a mere set-point regulator that fails a barrier from a regenerator that routes around it. The agentoscope does not detect a soul. It reports where, at what scale, and under which interventions a goal is the better explanation, and its honest use points it at organoids, wounds, and regenerating tissue rather than at the sun. ## 1. The Instrument Already Exists

Every optical instrument had a biological prototype. The lens that makes a microscope was already at work in the eye before anyone ground glass, which is why the microscope was discoverable: the function existed, implemented in tissue, and the instrument extracted and enlarged it. The idea behind an agentoscope is that agency-detection is like this. Humans already run a detector, well enough to see agency in a baby, a tiger, a dog, an adversary, and badly enough to see it in a thunderstorm or a sunspot or two triangles chasing each other across a screen. The competence is real and implemented. So the question is not whether agency can be detected, since we do it

constantly. The question is what the detector computes, whether it can be lifted out of the human case and calibrated, and whether it can then be pointed at systems we do not ordinarily read as agents, a slime mold, an organoid, a regenerating liver.

The first move is to say what the instrument must not be. It cannot be a detector of agency as a substance, a hidden essence some things have and others lack, because there is nothing to point it at and no way to be wrong in a useful direction. What it detects instead is a modeling fact: whether a system's behavior becomes more compressible and more controllable when we treat it as pursuing, preserving, or regulating something, rather than as merely being pushed around. Agency, on this reading, is not seen. It is inferred, from patterns of pursuit, correction, and self-maintenance, as the better of two competing explanations.

That reframing is what makes the project tractable, and it also produces the paper's main surprise. If agency is the victory of a goal-model over a passive model, then detecting it requires the two models actually to disagree, and the central fact developed below is that under ordinary observation they usually do not. A system quietly holding its state looks the same whether it is falling into an attractor or defending a set-point. The disagreement, and therefore the agency, shows up only when the system is perturbed in the right way. The agentoscope is less a camera than a prod.

2. What the Human Detector Is Doing

The human instrument is not one faculty but a stack of cues, and cognitive science has named most of the layers. At the bottom is animacy perception, the reading of self-propelled and contingent motion as alive, demonstrated in its purest form by Heider and Simmel, who showed that people watching simple shapes move will narrate a story of chasing, hiding, and bullying, complete with motives, from nothing but trajectories (Heider and Simmel, 1944). Above it sits the perception of causality, Michotte's finding that when one object's motion reliably launches another's, we see the causing directly rather than infer it (Michotte, 1963). Above that is goal attribution, the reading of behavior as efficient toward an end, which infants already do in the first year, expecting an agent to take the shorter path to its target and being surprised when it does not (Gergely et al., 1995; Gergely and Csibra, 2003). Formal cognitive science has modeled this top layer as inverse planning, recovering an agent's goals by inverting a model of how goals produce rational action (Baker, Saxe, and Tenenbaum, 2009), which is the same structure a philosopher describes when treating agency as the intentional stance, a predictive strategy of ascribing beliefs and desires rather than a discovery of inner stuff (Dennett, 1987).

Two things about the human detector matter for building a better one. The first is that its deepest cue is not motion but future-directedness: what makes a trajectory read as agentic is that it seems governed by where the system is trying to go rather than only by where it has been. The second is that the detector is miscalibrated, and known to be. Under ambiguity, threat, or mere noise, people over-attribute, seeing faces in clouds and intentions in weather, a bias documented as a hyperactive agency detection device and argued to be the cognitive root of much supernatural belief (Barrett, 2000; Guthrie, 1993). An instrument built to extract the human competence must therefore also correct the human error, and the correction, as it turns out, is exactly the operation

the passive versus agentic comparison forces.

3. Agency as a Model Comparison

Make the instrument precise. Let a system produce a trajectory of states, and compare two models of it. A passive model explains the trajectory through autonomous dynamics: forces, relaxation toward equilibrium, noise, whatever moves the system without reference to any goal. An agent model explains the same trajectory as the pursuit of a goal, a policy that reduces the gap between the current state and a target the system is trying to occupy. The agency score is the log-likelihood ratio of the two,

$$A(S) = \log \frac{P(\text{data} \mid \text{agent model})}{P(\text{data} \mid \text{passive model})},$$

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). The instrument does not announce that a system has a soul. It reports that the system's behavior is better compressed by a model containing a latent goal, and it reports how much better, which is a quantity one can argue about with evidence.

This is worth stating carefully, because the word agency carries more than the instrument measures, and the extra weight is where the project could go wrong. Agency in this sense is neither intelligence nor consciousness nor personhood. A system can score high because it defends a setpoint without being able to solve novel problems, which would be intelligence; without there being anything it is like to be it, which would be consciousness; and without any claim on moral or legal standing, which would be personhood (Barandiaran, Di Paolo, and Rohde, 2009). A thermostat has a little agency and no intelligence. A liver may have organ-level agency and no experience. Keeping these apart is what saves the agentoscope from panpsychist mush: it is a meter for one specific thing, the degree to which a goal is the better model, and it is silent on the others by design.

4. Why You Cannot See It

The reason agency is hard to detect is not that the signal is faint. It is that under observation the signal is not there at all, and this can be stated as a small theorem rather than a worry.

Consider the cleanest case. A goal-seeker that moves to reduce the distance between itself and a fixed target follows a trajectory that relaxes toward that target and then holds. But a passive point attractor, a ball settling in a bowl centered on the same spot, follows a trajectory of the same mathematical form, linear relaxation toward a center. When the goal does not move, the goal-model's dynamics are a special case of the passive model's dynamics, which means the passive model can fit everything the goal-model can fit, and fit it at least as well. The agency score cannot be positive. In a simulation of three hundred goal-seekers and three hundred passive attractors observed at rest, the detector's separation between the two populations is an area under the ROC of 0.49, indistinguishable from chance, and across all three hundred agents the score never rises

above zero, peaking at negative 0.06. Watching does not work, and not because the watcher is bad. It cannot work.

There is a deep reason, and it is the good-regulator theorem: any system that effectively regulates some variable must contain a model of the thing it regulates, so a well-run controller and the world it controls come to mirror each other (Conant and Ashby, 1970). But that mirror is internal. From outside, a system perfectly holding a variable steady and a system in which that variable is simply at rest present the identical face. The regulation is a fact about what the system would do if the variable were pushed, and if nothing pushes it, the regulation is invisible. This is why animacy perception leans so hard on motion and contingency, and why still or steady systems, however alive, do not trigger it. The agency is real and the observation is blind to it.

5. Why You Must Perturb

If agency is a fact about what a system would do under disturbance, then detecting it is an interventional problem, not an observational one, and it lands squarely on the higher rungs of Pearl's ladder of causation, where questions are settled by doing rather than seeing (Pearl, 2009). The operation the agentoscope depends on is the deliberate perturbation, the do-operation applied to the system's putative goal or its path, and the reading of the response. This is the modern form of a very old idea, that purpose is behavior organized around a goal by negative feedback, and that you demonstrate it by disturbing the system and watching it correct (Rosenblueth, Wiener, and Bigelow, 1943).

The simulation makes the reversal concrete. Take the same goal-seekers and passive attractors that were indistinguishable at rest, and now intervene: move the goal to a new location and see whether the system follows. The goal-seeker tracks the moved target, a motion only the goal-model, which has access to the goal variable, can predict, while the passive model, which is autonomous, is blindsided by every jump. The score climbs by roughly 101 nats per intervention, and the detector that scored at chance under observation now separates the two populations at an area under the ROC of 1.0. The agents run to a mean score in the hundreds; the passive systems, on which the same goal-perturbations are performed and simply ignored, sit near or below zero. Agency was there the whole time, and it took a perturbation to make it a fact about the data.

One subtlety decides which perturbations are worth performing, and it matters both practically and conceptually. Not every disturbance informs. If you merely shove a system away from where it sits, both a passive attractor and a goal-seeker return, so the recovery is shared and the score barely moves, staying near negative 2.3 in the simulation, no better than observation. The perturbations that reveal agency 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 trajectory. Mere resilience is not agency, because a rock rolled uphill also returns. The signature is narrower and is best stated as counterfactual goal preservation: the system keeps reaching or holding the same target across changes of path, obstacle, and starting point that a passive process could not survive.

6. Richness Is a False Friend

The human detector's characteristic error is to read agency into complexity, to see the storm's churn or the fire's flicker as striving. The agentoscope must not inherit this, and the model comparison is what protects it, because complexity and agency turn out to be different measurements that happen to co-occur in the animals we know best.

Run the instrument over a battery of systems and score each for agency under intervention and, separately, for dynamical complexity, the raw unpredictability of its trajectory to the best passive model. Across this battery the two scores run against each other, at a correlation of negative 0.63: the richer a system's dynamics, the lower its agency score tends to fall. A chaotic flow, all vortices and sensitive dependence, has the highest complexity in the battery and the lowest agency score of any system, around negative 527, because when its putative goal is moved it does nothing goal-like and the goal-model is punished for expecting it to. A chemotactic cell, whose trajectory is simple almost to the point of dull, tracks its moved target and scores around positive 550. This is the quantitative form of the intuition that a hurricane is a false positive and a liver is a real candidate. The hurricane maintains structure and displays gorgeous dynamics, but it preserves no internal variable against intervention. The liver holds its regime, corrects its outputs, and rebuilds toward a target when injured, and would answer a perturbation the way an agent answers it. Richness is not the signal. Answering is.

7. The Ladder and Equifinality

Agency, measured this way, is not a yes or no. It is graded, and the instrument orders systems along a ladder. At the bottom, scoring below zero, are the passive systems: a diffusion that goes nowhere in particular, a gradient that settles into its basin, a chaotic flow that ignores every intervention. Above zero, in ascending order, are the goal-directed ones: a thermostat that holds one variable, a chemotactic tracker that pursues a moving target, a regenerator that pursues a target and routes around obstacles to reach it. The ladder is not imposed. It falls out of how strongly each system's behavior is coupled to a goal it defends under perturbation.

The top of the ladder needs its own criterion, because tracking a goal is not the strongest thing an agent can do. The stronger property is equifinality, reaching the same end by different means, which von Bertalanffy identified as the mark of an open system pursuing a goal rather than merely running a fixed process (von Bertalanffy, 1968). Test it by putting a wall between a system and its target. A plain set-point regulator, a controller that only ever moves straight down the gradient toward its goal, is stopped by the wall and never arrives, reaching the blocked target in none of a hundred trials. A regenerator that treats the goal as an end to be reached by whatever path routes around the wall and arrives every time. On the agency score under simple goal-shifts the two look nearly equal, both plainly agentic. Only the obstruction pulls them apart, and it pulls them all the way apart, zero against one. This is the operational heart of the liver and the embryo cases that motivate the project, where a disturbed system does not merely resist but navigates to a target configuration by a route it was never going to take, which is the behavior the basal-cognition

literature reads as goal-directedness in anatomical and physiological space rather than in the space of movements (Levin, 2019; Friston et al., 2015). It is also the reason the agentoscope's best frontier is downward and inward, into tissue and morphogenesis, rather than outward into the weather that so reliably fools the unaided eye.

8. What the Agentoscope Cannot Certify

An instrument is trustworthy in proportion to the clarity about what it does not measure, and this one has real limits that its numbers can hide.

It is relative to the models compared, which is the deepest limitation and not a removable one. The agency score is high or low only against a specific passive alternative, and a sufficiently clever passive model can always be built to absorb a given behavior, at the cost of complexity and lost generality. What the instrument reports is therefore not agency in the absolute but the margin by which a goal-model out-predicts the best passive model one has thought to write down, which is why the honest output is a score against a named alternative and a scale, never a verdict of agent or not. The good-regulator theorem cuts here too: since a good enough passive model of a regulated system must resemble the agent's own model, the two families converge in the limit, and the instrument measures how far short of that limit the passive account currently falls.

It requires a boundary it does not itself supply. Before scoring anything the instrument must be told what the candidate unit is, where the system stops and the environment begins, and that individuation is a genuine and unsolved problem, whether posed as drawing a Markov blanket around a self-organizing region or as identifying the self-producing closure of an autopoietic system (Friston, 2013; Maturana and Varela, 1980). Choose the boundary wrong and a regulating organ can look like passive plumbing, or a passive aggregate can look like a striving whole. The agentoscope inherits this choice rather than making it.

It can be fooled, in both directions, and its cure has a cost. The human detector over-attributes, and a naive agentoscope run in observation mode would inherit the bias, since without intervention the score is uninformative and any decision threshold placed on it produces false positives. Intervention is what collapses the over-attribution, which is the instrument's genuine advance over the human prototype and over passive statistical agency-detection alike. But intervention is not always available. You cannot move the sun's goal to see whether it tracks, and for systems that cannot be perturbed the instrument is silent, which is the correct answer where the naive detector would have guessed. There are also other operationalizations of the same underlying idea, empowerment as an agent's control over its own future sensory states, or active inference as the minimization of expected surprise, and the model-comparison score developed here is one member of that family rather than its final form (Klyubin, Polani, and Nehaniv, 2005; Friston, 2010).

None of this dissolves the project, and stating the limits sharpens what remains. The agentoscope does not find souls, settle consciousness, or confer personhood. It takes a competence humans run automatically and unreliably, the reading of goals into motion, and turns it into a procedure with a number and a scale, one that refuses to score until it has perturbed, that scores complexity and

striving separately so that storms stop passing as minds, and that grades agency along a ladder from a settling gradient to a regenerating tissue. Pointed at the systems it was built for, not the weather that fools the eye but the organoid, the wound, and the regenerating organ, it measures the one thing it can honestly measure: where, at what scale, and under which interventions a goal becomes the better account of what a piece of the world is doing.

References

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.

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.

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

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface*, 10(86), 20130475.

Friston, K., Levin, M., Sengupta, B., and Pezzulo, G. (2015). Knowing one's place: a free-energy approach to pattern regulation. *Journal of the Royal Society Interface*, 12(105), 20141383.

Gergely, G., and Csibra, G. (2003). Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7), 287–292.

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

Guthrie, S. E. (1993). *Faces in the Clouds: A New Theory of Religion*. Oxford University Press.

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

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

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* (pp. 128–135). IEEE.

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

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

Michotte, A. (1963). *The Perception of Causality* (T. R. Miles and E. Miles, Trans.). Basic Books. (Original work published 1946.)

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

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

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