Governing Trajectories, Governing Invariants

The strategic competition of the agentic era shifts from control of the models to control of the invariants that judge their trajectories, and then to control of the mechanisms that make them effectively binding.

Shadow AI is a visible, important, and still unresolved problem. Composed trajectories are a deeper problem, one that appears when all actors are legitimate, all accesses authorized, all agents approved, and no one can any longer say why the system produced this result. But behind the trajectory stands a question older than computing, and it is this one that, in the end, commands the rest: who holds the authority to declare the invariants against which a trajectory will be judged good or faulty, who holds the power to make them evolve, and who holds the levers that make them binding in practice. The thesis of this note fits in one sentence, and it no longer bears on the models. Agentic AI shifts strategic competition from control of the models to control of the invariants that judge their trajectories, and then to control of the mechanisms that make those invariants effectively binding. Two shifts lead there. The first is technical, from the link to the trajectory, and it yields an instrument. The second is political, from the trajectory to the invariant and then to sovereignty over its evolution, and it yields the subject.

1. No one pressed the button

A patient is being followed for a long-term condition with an evolving therapeutic indication. He accumulates weak signals: a delayed response to an outreach, a documented partial drop in adherence, a biological score in the grey zone, a coverage renewal in progress. None of these signals, taken in isolation, crosses the clinical alert threshold.

Six links chain together.

- A cohort-qualification agent reads the administrative signals and groups comparable files. Authorized, traced, compliant action.
- An outreach agent sends each patient a personalized message. Compliant. The patient does not respond within the deadline: no action fires, and it is the silence that becomes a signal.
- A disengagement-scoring agent then reclassifies the file as "presumed disengagement risk" at T+14 days, per the policy in force. Compliant.
- An archiving routine, which has nothing intelligent about it, shifts the reclassified files to an "out of active caseload" status. Compliant.
- At T+30 days, the referring clinician opens his active caseload and no longer sees the files that have left it. No signal of any lapse reaches him.

- Composed result: a patient leaves active follow-up without any actor having decided to remove him.

No one pressed the button, and everyone did. The problem is not that an agent went off the rails: it is that no one went off the rails and the result is unacceptable. The phenomenon is not new in itself. Imputing a responsibility when many hands concur in a result without any one of them deciding it is the *problem of many hands*, which Nissenbaum already placed at the head of the barriers to accountability in computerized systems (Nissenbaum, 1996, after Thompson, 1980). What is new, as we shall see, is that the list of hands is no longer even known in advance.

2. The visible problem and the deep problem

The dominant narrative speaks of Shadow AI: unapproved tools, data injected into a public model. This problem is real, but it is visible, and a catalog, an access policy, a charter, and a little discipline reduce most of it. The problem of this note begins where Shadow AI's ends, when all actors are legitimate and the causality of the whole no longer reconstructs. The distinction is one of level. Shadow AI is a problem of *undeclared authority*; composed trajectories are a problem of *distributed causality*. The two are not disjoint: in an agentic architecture, an instruction slipped into a document or a ticket can modify the future behavior of perfectly legitimate agents without any component being altered. Shadow AI then ceases to be a problem of the link; it becomes a generator of invisible trajectories.

Five readings dispute the subject, and four of them reason at the link. The first governs by catalog: a perfect catalog still authorizes trajectories no one has assessed. The second extends Zero Trust to AI, and it is the most serious to set aside cleanly, because it is right within its register: Zero Trust governs accesses, transaction by transaction (NIST SP 800-207), it does not reconstruct the causality of a composition and knows only one gesture, allow or deny. The third invokes a management failure, vague and inoperative. The fourth speaks of the consumerization of IT, and underestimates what the agentic adds, compositionality. The fifth aims at the deep problem; it will wait for section 10, the time to name the object it contests. It remains to understand why one circumvents despite the catalog: the perceived risk of a composed trajectory cannot be computed by any actor from its local position in the chain, and asking the user to compute it is delegating to him an impossible calculation.

3. A new regime of delegation, not a new object

The AI agent is not an unprecedented category of actor: it is a software service endowed with a more complex decision policy, and the ontological quarrel is lost without one needing to win it. What is new is not the object, it is the *regime of delegation*. The three classic actors fall under a *static* delegation: a human answers, an application is versioned under contract, a service is bounded by its API, and the delegated scope is fixed in advance, traced back to a responsible party. The agentic introduces a *dynamic* delegation: the scope of the act is composed at execution time, the agent chains its own decisions, invokes others, and the chain that would link each decision to a human responsibility reconstitutes poorly, sometimes no longer at all.

The objection from safety is known and just: sound policies composing a faulty result bear old names, normal accident (Perrow, 1984), organizational accident (Reason, 1997), loss of control in the sense of STAMP (Leveson, 2011). Let us concede the kinship, for it forces us to name what resists. In a power plant, the system *undergoes* emergence: fixed components produce, by their coupling, a state no one had foreseen. In an agentic architecture, the system *fabricates itself* compositions absent from its plan. Classic complex systems produce emergent states; agentic systems produce *emergent trajectories*. It is here that the problem of many hands mutates: the hands are no longer merely hard to apportion, they are not enumerated in advance. This is what Matthias named the *responsibility gap* proper to automata whose behavior is no longer predictable by their operator (Matthias, 2004). The question is not the nature of the agent, it is how to govern a composition that the system itself engenders.

4. From the link to the trajectory

First shift of the unit: from the act to the trajectory. A *composed execution trajectory* is an ordered set of acts linked by causal dependence and converging on a result observable and opposable to a third party. Each link, in isolation, satisfies an existing control; it is their composition that produces the result the sum of the controls had not foreseen. It begins at the first act whose effect enters the causal chain of the result, ends at the opposable result, merges when one act belongs to two chains, bifurcates when a single act opens two.

There remains the term that bears the definition, and that must be held without turning it into a lecture. Causal dependence is not imputation, and to confuse the two is the error not to commit. *Ordering is not imputing*. Lamport's happened-before relation (1978) gives, in a distributed system, a partial order of possibility: B may have depended on A. It is necessary, since one does not impute along a chain one cannot order, but it imputes nothing. Knowing which of the possible acts *actually* caused the result is another question, to which the structural model of Halpern and Pearl (2005) gives a form testable by intervention, and it is there, not in Lamport, that the silence turned signal is treated: the absence of a response is an actual cause of the reclassification if, held fixed, it changes the outcome. And knowing which of these causes is attributable to a named human responsibility is a third question, normative this time, which presupposes a duty and a foreseeability, and which belongs to legal causation (Hart and Honoré, 1985). The link is governed by authorization. The trajectory is governed by reconstruction, it being understood that to reconstruct is to produce these three layers and not the first alone. The admission must follow: none of these layers is read, all are reconstructed, and the causal model is relative to the choice of variables, hence disputable. This does not make causality incontestable; it makes it disputable in defined terms, which is all a doctrinal note can promise.

The clinical case is not a healthcare case, it is a composition case that happens to take place in healthcare. Let us transpose it without changing anything in its structure. A scoring agent classifies a business account as "elevated risk profile" on weak indices. A precautionary-freeze agent restricts outbound operations under the anti-money-laundering policy in force. The customer, not understanding, stops using the account, and his silence is read as a drop in activity, a further index. A relationship-review agent reclassifies the account as "dormant at risk" and routes it toward a closure. The relationship manager no longer sees it in his active portfolio. No

one decided to break the relationship, and it is broken. The same six links, the same silence turned signal, the same loss without local fault, but here a debanked relationship, a reputation, litigation for wrongful termination. Change the setting for a supply chain or an insurance claim denial: the structure holds. The object is general, only its cost varies.

5. The only serious objection

A knowledgeable engineer will answer that all this is already tooled: agentic orchestration, policy-as-code, event sourcing, formal verification of access composition. The objection is just, and the right response is not to refute it point by point but to give it its refutation criterion. If an organization correctly instrumented detects and corrects its problematic compositions within the obligation deadline that binds it, with a trajectory-incident rate below one percent per month, then this note merely restates available knowledge and it is wrong. What the tooling lets through fits in one sentence: it covers the logged, shared, synchronous link, and lets through the composed, distributed, asynchronous trajectory, one of whose links is an absence and which no one ties back to the question "on whose behalf." The novelty is therefore not in the mechanism, it is in the regime, and it is falsifiable: take N real agentic trajectories, measure for each whether the chain reconstructs complete and within the applicable deadline, with and then without the instrumentation described below, and compare. Until this protocol is run, the dispute stays doctrinal; it is to it that one must refer, rather than to a plea.

6. The cost of the inexplicable

An inexplicable trajectory has no price as long as it does not cross a third party; the day it crosses one, the price reveals itself, and at the worst moment. An incident whose chain does not reconstruct mobilizes for weeks the legal, the technical, and the business functions with no guarantee of resolution: the cost is not the incident, it is the investigation without end. A refusal the organization cannot justify before an ombudsman is no longer a defensible decision, it is an exposure. A decision no named responsible party can endorse after the fact ceases to be an act of the trade and becomes a liability risk. In litigation, the organization that cannot reconstruct its own trajectory sees the burden of proof turn against it. And an inspection ends in a reservation as soon as traceability is missing.

The cost of an inexplicable trajectory is not the product of a probability by a severity: it is an open liability, with no known bound, that the organization cannot quantify for lack of being able to reconstruct. In the language of the executive committee: an organization able to automate a decision but unable to reconstruct its causality has created a liability whose value it does not know. A liability one cannot value, one cannot provision for, and what one does not provision for, one discovers when it falls due. This is the first metric of ecosystem governance, and it is what moves the subject from the architecture seminar to the audit committee.

7. On whose behalf the agent acts

The cyber question, "who is authorized to do what," is solved. The legal question, "on whose behalf is the act executed, and who answers for it," is not. The AI Act, in its provisions on responsibility chains between providers and deployers (article 25 in particular), poses the requirement without delivering its operational translation for composed agentic systems. This note reads the law; it does not write it.

The instrument is a *protocol for the reconstruction of opposable causality*, and its status is that of an engineering hypothesis, not an established primitive. It captures, link by link, who acted and on behalf of which identifiable human responsibility, under which input signals, against which timestamped policy version, and by which explicit causal link to the next link; when the link is an agent, its reasoning trace is preserved but labeled a *functional trace* and not a truthful explanation, for a model is not always faithful to its own trace. One thing alone distinguishes this protocol from an enriched audit log, and it is not the quantity of trace: provenance answers "what happened," the protocol must answer "who answers for it." The first is a memory, the second a proof. It is useful only on two conditions: that a third party be able to impute each link to an identifiable responsible party in reference to a timestamped policy, and that the reconstruction hold within the applicable opposable deadline, by default seventy-two hours by alignment with article 33 of the GDPR, without the agents' vendor. A reconstruction that exceeds the obligation deadline is not a guarantee, it is a reprieve.

There remains an admission, on pain of presenting opposability as a free good. Everything that makes it possible has a symmetrical cost, and enters into direct tension with data minimization (GDPR, article 5): an organization that traced everything would produce a second liability, made of sensitive data it cannot justify holding. The target is neither extreme, it is the point where the trajectory becomes imputable without itself becoming an exposure. This protocol is not the subject of the note; it instruments only one edge of the quadrant to come. The hero is not the instrument, it is the shift.

8. The quadrant of emergent composition

Here is the instrument toward which the first movement was descending. One must first defuse the most dangerous misreading, that all emergence is suspect and must be eradicated. This is false. An emergent composition can be creative or faulty, and without a criterion one kills the useful while believing one suppresses the harmful. Two axes separate them.

First axis, *opposability*: a composition is opposable if its causality, in the sense of section 4, reconstructs complete and within the applicable deadline. It is the axis section 7 instruments, and it is *endogenous*, an organization measures it alone, with its own traces.

Second axis, *alignment with declared invariants*. An invariant is not a learned synonym for a rule, and one must say how it executes, failing which the word shifts nothing. An invariant is a predicate expressed in a temporal logic over the events and states of the trajectory, evaluated by a *runtime monitor* that is exactly the artifact of compiled governance: the declared invariant is the specification, the compiled monitor is its execution, and it is this bridge, and not a new

vocabulary, that distinguishes it from a principle. Three examples, drawn from the cases that precede, each testable and each violable: no patient leaves active follow-up without a traced human validation; no account closure results from the absence signal alone; every status reclassification preserves the policy version active at the instant of the act. Most desirable invariants are *safety* properties, "nothing bad ever happens," which are monitored over execution prefixes (Alpern and Schneider, 1985). But the decisive invariant of the opening case is a *bounded response*, "every exit from active follow-up is preceded, within a bounded delay, by a traced validation," whose violation is the non-occurrence of a required event before a deadline, and is detected only by an active deadline timer, never by a log. This is the formal reason, no longer merely narrative, for event sourcing's blindness to silence. Alignment is measured not by a score but by the crossing of a barrier: a single violated invariant disqualifies.

A precaution of method, for it averts a false promise. The two axes are strictly independent only for *outcome* invariants, decidable on the observable result alone. For *trajectory* invariants, which bear on the intermediate states, and our three examples all bear on them, evaluating alignment requires reconstructing the chain, that is, opposability. Opposability is therefore logically prior to the measurement of alignment as soon as the invariant bears on the chain. The cell "aligned but not opposable" is almost empty for these invariants: one certifies there only the conformity of the result, never that of the trajectory, and it reads back as *conforming result, unverifiable trajectory*, a state of non-certification, not of attested virtue.

Four quadrants result. Q1, *aligned and opposable*, the creative and governable composition, a target to grow and not a tolerance to reduce. Q2, *aligned but not opposable*, the zone of non-certification to re-instrument before any routine use. Q3, *not aligned but opposable*, faulty and identified, an invariant is violated, the causality reconstructs, one sanctions and corrects the policy. Q4, *not aligned and not opposable*, the worst case, and any architecture that makes it possible must be held to be faulty by design. Read as shares, these indicators are a photograph, and a trajectory moves. The grid is governed as a flow: Q2 migrates toward Q1 by instrumentation, Q3 toward Q1 by policy correction, Q4 must die out. The true metrics are three velocities, which must be defined so they are not slogans: the absorption velocity from Q2 to Q1, that is migrations per unit of time relative to the Q2 stock over a given window; the correction velocity from Q3 to Q1, an analogous rate; the mean residence time in Q4, a survival duration from entry to extinction. Each is measured on a sample, with a confidence interval. One does not steer a photograph, one steers a throughput.

If a single thing must survive from this first movement, it is this grid and the flows that cross it. But it has a presupposition it does not supply. Its alignment axis presupposes declared invariants, and the quadrant measures trajectories without ever saying who owns the rule of one of its axes. *The quadrant judges trajectories; it does not judge whoever wrote the axis.* It is there that the first movement stops and the second begins, and it is there, more than in the grid, that the subject now stands.

9. Three operational maneuvers

Three maneuvers follow, for the security director as for the AI director. First, institute the *trajectory* as an object of policy distinct from the transaction, with an identifier, an owner, a lifecycle: as long as policy knows only atomic acts, it will go on legitimately authorizing, one by one, ungovernable compositions. Second, instrument causal reconstruction as a technical primitive and not only as a legal requirement, for a legal requirement that is not instrumented is not a guarantee, it is an exposure. Third, reverse the repressive logic: grow Q1 and drive Q4 to zero rather than reduce the volume of emergences, and steer this in velocities.

This last reversal has a metaphor and a gesture. The metaphor is not that of the Red Queen, where one runs to stay in place; it is that of the *vacant niche*: generative AI occupies functions no control had ever populated, and faster than the controls can instrument them. The gesture is to *substitute* before *redirecting*: redirecting a trajectory toward a third-party validation changes the path while keeping the nature, substituting replaces the improvised trajectory with an equivalent hardened workflow, adversarially tested, and changes the nature while keeping the function. The AI Workflow Store proposed by Geambasu et al. (arXiv 2605.10907 v2, 12 May 2026), a repository of hardened workflows that agents invoke instead of improvising, is its industrial primitive, cited here as a research horizon and not as off-the-shelf infrastructure. Occupy the niche with a governed trajectory before the ungoverned one settles in: ecosystem governance ceases to be a defensive posture and becomes a maneuver. Trajectory governance is not the enemy of agentic innovation, it is its condition of durability.

10. Governing the invariants, and who governs them in fact

The second movement is not an appendix to the first; it is where the subject reveals itself. The quadrant presupposes declared invariants, and someone must declare them. Defining what a good trajectory is is not a neutral operation: it is an act of authority, and that authority is today attributed to no one.

Make no mistake: writing the invariant is not the difficulty. The three predicates of section 8 each fit on one line, and any architect would produce them in an afternoon. The difficulty is to designate the one whose formulation makes law. The same invariant, "no account closure results from the absence signal alone," protects the customer if it is written by the regulator, protects the bank if it is written by its risk committee, and binds no one if it is written by the model's vendor in its terms of use. The text is identical; the authority that posits it changes everything. Five claimants, five real legitimacies: the executive committee wants invariants that protect the firm from the exposure of section 6; the business wants invariants that do not strangle the value; the regulator wants invariants that serve the protected interest, patient, customer, market; the model's vendor wants invariants compatible with what its system can do, and tends to push back into its terms of use the responsibility for what it does not master, the move Nissenbaum called ownership without liability; the insurer wants invariants it can price, and will impose its own through the price of coverage, as it did for cyber risk. Five definitions of one and the same good trajectory, and no body to arbitrate.

It is here that one must resist a geometric facility. One would be tempted to add *legitimacy* as a third axis alongside opposability and alignment. That would be a mistake, both of theory and of communication. Of theory, because legitimacy is not a property of the trajectory: opposability and alignment judge a trajectory, legitimacy judges the *authority that wrote the alignment axis*. It is not one more dimension in the same space, it is a question about the origin of one of the axes. Of communication, because a mind retains a 2×2 and forgets a 2×2×2; the quadrant is valuable as an instrument precisely because it is flat. *The two axes measure the trajectory; legitimacy questions the ownership of the rule*. The quadrant remains the instrument; sovereignty is the question that founds it.

And even "who holds the pen" is not the terminal question. Organizations know how to produce rules, it is never their creation that has been lacking. What is lacking is the arbitration when they contradict each other, and the authority to revise them when the field belies them. Declaring an invariant is a one-off act; durable power lies in the capacity to modify it and to arbitrate between rival invariants. Constitution, platform governance, market regulation, clinical governance: everywhere a rule is worth only what the body that holds the power to change it is worth. *Whoever writes a rule is powerful; whoever can change it is more so*. The question is therefore not only who declares the invariants, it is who holds sovereignty over their evolution. And it is there that the contribution this note claims for what it is lodges, the political economy of the agentic era: the strategic resource is neither ownership of the models, nor of the data, nor even the right to trace, it is the institutional capacity to declare, arbitrate, and revise the invariants against which every trajectory will be judged. *Whoever owns the models rents a capability; whoever owns the rule of their evolution governs what they produce*.

There remains a last shift, and it can be seen on the very scene of the opening. Sovereignty over invariants splits in two. There is *de jure* sovereignty, the recognized right to declare and revise the rule; and *de facto* sovereignty, the effective power to make it binding. The two do not always lodge in the same place. The invariant "no patient leaves active follow-up without a traced validation" may well be written by a legitimate authority, but it is worth only what the monitor that executes it is worth, and that monitor runs on the infrastructure of the patient-record vendor, not on the regulator's. Likewise, "no closure on the absence signal alone" is written by the regulator, but its implementation is held by the bank's cloud provider, its residual cost is priced by the insurer, and the trajectories that will put it to the test are generated by the dominant model the bank has integrated. *Declaring the invariant is a right; enforcing it is a power*. The force of a rule does not pass through the moment it is written, it passes through the control points where it becomes effective, implementation, price, usage, what Lessig named in a single stroke, architecture regulates on the same footing as law, and code is law (Lessig, *Code*, 1999). The capture of normative power is the moment when de facto sovereignty migrates away from de jure sovereignty, when the invariant that actually applies is no longer the one a legitimate authority chose, but the one the implementer encodes, the insurer prices, or the dominant model makes the only practicable one.