# You Govern Only What You Can Still Redirect

## Governability is not inventoried; it is tested under perturbation, and it is distinct from resilience.

## 1. What decides is not the presence of control, but its performance under perturbation

When Arizona requires, effective 1 July 2026, that a medical director personally sign any denial of coverage produced by an algorithm (HB 2175), and when Colorado does the same as of 30 June (SB 24-205), the legislator obtains exactly what it aims for: accountability.

A signature designates a responsible party, opens a remedy, grounds a challenge. These are legitimate objectives, and the signature serves them. What it does not serve is the governability of the system. It attests that a human was present; it attests nothing about the system's capacity to catch the decision if it turns out to be wrong three weeks later, when the patient has already left the circuit. The signature is necessary to accountability. It is insufficient for governability. The two are not the same.

This is the blind spot of the discourse on the human in the loop. We assess a supervision arrangement by its capacity to correct the errors we expected. We almost never assess it on what actually decides the fate of systems: their behavior in the face of what they had not anticipated. Known errors fill the everyday. Unforeseen perturbations redraw systems.

The thesis of this note holds in one sentence, and it is more general than any particular mechanism: the governability of a system is not demonstrated by the presence of safety mechanisms, but by their performance when they are put to the test. A circuit breaker that exists without ever having been tripped is not a guarantee, it is a decoration. From this thesis follows a concrete condition, recovery capacity (which, as we shall see, takes two forms) and which is to be confused neither with the governability it conditions, nor with the resilience it subsumes.

Domain of validity, and a warning of level. The demonstration bears on automated systems endowed with an identifiable supervisory apparatus. It does not claim to cover distributed orders without a central supervisor (markets, digital commons, open ecosystems…), which fall under a different analysis. Above all, it borrows part of its counterexamples from aviation, electrical grids, crisis management. This displacement, from algorithmic supervision toward critical sociotechnical systems, is not an inadvertent

slide: it is claimed. These systems are the mature instances of the same relation of supervision under perturbation, and it is in that capacity, and that capacity only, that they serve as a test bench. The note therefore speaks of both at once, AI supervision and complex systems, but of a single object: the relation between an automated system and what governs it. Where the argument holds for the second without holding for the first, this will be said.

## 2. Three words that get confused: resilience, recovery, governability

The word that carries the demonstration must be defined before being used, on pain of becoming a semantic attractor that absorbs every problem. But an isolated definition does not suffice: the risk here is not the vagueness of one term, it is the confusion of three. Let us posit them as a chain, from the system toward its supervisor.

Resilience is a property of the system: its capacity to absorb a perturbation without external intervention (redundancy, containment, graceful degradation, margins...). A resilient system takes the hit alone.

Recovery is an operational property, and it is the one that must receive an observable criterion lest it become the new catch-all. I call it the capacity to restore an acceptable space of states after the crossing of an unrepresented perturbation. Three objects are named there and therefore measurable:

1. a space of states judged acceptable,
2. an event that exits it,
3. a return into the acceptable domain.

Recovery comes in two forms, we will get to them; but it is not "everything that helps a system survive." It is this precise movement: to restore, after exit, a defined domain. For example, the activation of a Disaster Recovery Plan (DRP) after a disaster (an event that exits the space of states judged acceptable).

Governability, finally, is not a property of the system. It is a property of the relation between the system and its supervisor: the latter's capacity to maintain the system within an acceptable space, or to bring it back there, attested under the test of perturbation and across regimes. The weight is on "under test." Governability is observed under load; it is not deduced from an inventory of mechanisms.

This chain settles at a stroke two objections that would otherwise ruin the note.

1. The first is the objection of tautology: if recovery encompasses rollback, resilience, containment, redundancy, degradation and repair, then "governable" ends up

meaning "capable of continuing to be governed," which is true and circular. The riposte is not to narrow recovery but to situate it: recovery is what the system makes available; governability is what the supervisor manages to mobilize of it, and what has been put to the test. A system can be highly recoverable and poorly governable, because the supervisor does not have its hand on its return levers, does not see them, or does not dare to actuate them. Recovery is necessary to governability; it is not its definition.

2. The second objection is more serious. A reader familiar with Perrow, Hollnagel or Woods will say that all of this reformulates the resilience of complex systems under a new vocabulary. The objection would be decisive if governability were a property of the system, for it would then be only a belated synonym for resilience. But it is a property of the system/supervisor relation, and that is exactly what the resilience literature leaves in shadow: a system can be intrinsically resilient and yet escape its supervisor, and a barely resilient system can remain governable because an informed supervisor can catch it in time. a. *Resilience describes what the system does alone.* b. *Governability describes what a third party can still do with it.* Resilience is a virtue of the system; governability is a virtue of the couple. It is this difference of object, and not a new word, that makes the contribution.

Having governance resources is, moreover, not the same thing as being governable. The French army of 1940 had observation, hierarchical authority, the means of intervention. It collapsed because these resources were not articulated to the rhythm of the event. Conversely, grid operators have held critical systems with rudimentary means, because their articulation was right. The causality "means, therefore governability" is false. What is measured is the holding of the means under constraint, not their number.

Three distinctions, finally, prevent confusing governability with its immediate neighbors.

1. It is not alignment: a perfectly governable system can serve a destructive end, it is neutral on the object.

2. It is not instantaneous pilotability: a pilot disengages the autopilot, he does not rewrite the control laws in flight; acting *in* the system and acting *on* the system are two distinct powers.

3. It is not local controllability: a system uncontrollable point by point can remain governable as a whole. But the separation from alignment, useful as it is, is not watertight: one controls only relative to a finality, and contradictory objectives render a system ungovernable however richly endowed, because the authority no longer knows in the name of what to intervene.

## 3. Recovery, and its two forms

If governability is a performance under perturbation, the property of the system that makes it possible is recovery capacity. The precision of §2 is essential here, because recovery has meaning only relative to a threshold and to variables. A denial of reimbursement may be legally reversible when the clinical harm it caused no longer is, biologically. What counts is not reversibility in general, it is the recovery of the variables that decide survival, and before the point of no return.

This capacity takes two forms, which must be separated because they are confused.

1. The first is reversibility: to go back, to undo. It is the most powerful form when available, and it itself comes in several registers that must not be held equivalent: technical reversibility (rollback of a system), decisional (annulment of a decision), legal (avenue of remedy), physical or clinical (repair of a real harm). A system can be reversible in one register and not in another, and that is precisely where the illusions of safety lodge: one certifies a technical rollback while leaving intact a physical harm already done.
2. The second form is resilience: surviving without going back. Many critical systems operate on largely irreversible decisions (surgery, aviation, nuclear crisis management, the conduct of war…). They remain governable nonetheless, not because they can go back, but because they absorb the harm. The correct formulation is therefore not that reversibility is the bedrock of everything, but that reversibility is the most powerful form of recovery, and that where it is impossible, resilience takes its place. Recovery capacity is necessary; reversibility is its favorable case, not its sole modality.

One will note the exact relation with §2: resilience appears here as one of the two forms recovery can take, on the system side, and not as a rival concept. What governability adds is the question that neither reversibility nor resilience poses of itself: are this return or this absorption within reach of a supervisor, at the right moment, and has it been verified?

## 4. Why some systems absorb surprises and others collapse

A demonstration resting only on catastrophes proves little. Three Mile Island, Challenger, the failure of Long-Term Capital Management, the Flash Crash of 2010, Fukushima: the list is striking, but it contains only systems where the perturbation won. One can draw up the symmetrical list, that of systems that absorb the unforeseen without collapsing: air traffic control absorbs out-of-procedure situations every day, electrical grids continuously manage local failures, modern civil aviation has made the unforeseen

incident a treated case rather than a fate. A serious theory must explain both lists, not only the first.

What separates the two lists is not luck, it is mobilizable recovery capacity. The systems that hold have built, upstream, the wherewithal to absorb or undo: redundancy of air traffic control, containment and load shedding of grids, margins and recovery procedures of aviation. The systems that collapse had, at the decisive moment, crossed an irreversible threshold without capacity for return or absorption. Catastrophes therefore do not refute the thesis, they instance it: they are the cases where recovery was missing.

The word around which everything turns must still be stabilized. "Surprise" is too vague: it designates now the improbable event, now the unknown, now the slow drift. The pertinent category is more precise: a perturbation not represented in the design assumptions. What counts is not that an event be rare or spectacular, but that it fall outside what the system planned to treat. A slow model drift and a brutal shock belong to the same category as soon as neither figured in the assumptions, and the first is often more dangerous than the second, because it triggers no alarm.

But this definition by non-representation comes at a difficulty that must be confronted, not skirted. An unrepresented perturbation identifies itself as such only after the fact: as long as it has not occurred, it is, by construction, outside the field of what we know how to name. How then to demonstrate ex ante that a system will know how to treat what, by definition, is not yet known? The honest answer is that one cannot, and that one must renounce claiming it. One never demonstrates the capacity to manage a specific unknown perturbation. One demonstrates only the presence, and the performance under test, of generic recovery capacities: width of the restorable state space, speed of return, independence of the catch-up paths, margin before the irreversible. These capacities are exercised on represented perturbations, which stand in as proxy for those that are not. The proxy is imperfect, and that is an irreducible limit, not a detail: exercising a recovery capacity on the known remains the only available datum on its behavior in the face of the unknown. Governability does not suppress uncertainty about the unprecedented; it displaces the bet, from "did we foresee this event?" toward "do we have a catch-up machine generic enough, and have we run it?" That is a better bet. It is not a proof of invulnerability.

## 5. Resources are a graph, not a pyramid

It would be convenient to arrange the conditions of recovery into a clean hierarchy. That would be too elegant. Recovery capacity rests on four resources that interact more than they stack:

1. observability (seeing what the system does),
2. intelligibility (understanding why),

3. authority (being able to act without asymmetric sanction),
4. intervention capacity (having the means to act).

Their dependence is not linear but crossed.

Without sufficient observability, recovery becomes unusable: one does not undo what one does not see, and a circuit breaker one does not know when to trip protects from nothing. But strong observability can compensate weak intelligibility: we pilot Watt's steam engine, antibiotics, deep learning well beyond what we explain of them, because we observe their effects and keep our hand on them. And without authority, intervention is fictional: a supervisor sanctioned for having blocked a flow learns to stop blocking.

The system resembles a graph of dependencies, not a pyramid, and one must resist the temptation to turn this heuristic into an ontology. One will note that these four resources are precisely the variables of the relation defined in §2: they describe not the system alone, but what a supervisor can see, understand, decide and do about it. That is why they fall under governability and not under resilience alone. The point is not to classify them, it is to put to a concrete system four opposable questions, plus a fifth that commands them all: what can it recover, and has it been put to the test?

## 6. Why real systems lose their recovery capacity

If recovery is the condition, how to explain that so many systems deprive themselves of it without anyone having decided so? It is here, and not in the taxonomy of resources, that the most fertile contribution lodges: passing from "what is necessary" to "why it disappears." §5 draws up the static inventory of what would be needed; the present paragraph gives the dynamic of its erosion. Four mechanisms together turn the stage set into a rational equilibrium, that is, a state toward which reasonable actors converge without anyone having willed it.

1. The first is economic: automation collapses the marginal cost of the decision, serious examination keeps its own. The thorough review, which cost little relative to a low volume, ceases to be profitable relative to a volume become massive. It is not rigor that becomes impossible, it is its cost/benefit ratio that inverts.
2. The second is organizational, and it is the most tenacious: responsibility is asymmetric. One sanctions the supervisor who blocked a legitimate flow. The error is visible, dated, attributable; one rarely sanctions the one who let a machine error through, for the fault dilutes in the system. The supervisor learns, without being told, that validating is risk-free and blocking carries one. He validates. It is optimal behavior for him, and destructive for recovery.
3. The third is cognitive: as the model goes on not erring, vigilance erodes by habit. Validation, at first a deliberate act, becomes reflex: one ratifies the system's

choices because it has never committed a gross error so far. The supervisor remains present, the signature of §1 is indeed there, but his presence has ceased to produce examination. This is automation complacency, and it is all the stronger as the model is good: a mediocre system would keep the guard up by its very failures.

4. The fourth is proper to the agentic, and it is the most recent: when a system itself orchestrates planning, delegation and execution, the chaining of decisions ceases to be legible. No one any longer knows why a given action took place, and the power to undo then bears on an object become opaque. One can retain authority and the means, and lose recovery nonetheless, for want of intelligibility of the path to retrace.

These four mechanisms do not add up, they reinforce one another: the economy suppresses the time for examination, the organization suppresses its incentive, cognition suppresses its habit, the agentic suppresses its object. Recovery does not disappear by decision, it erodes by dilution. None of these mechanisms presupposes an intention. It would therefore be wrong to write that AI turns control into theater, as if someone had set the stage. AI displaces two constraints, cost and intelligibility, and this displacement alone suffices to make the stage set the spontaneous equilibrium. The practical consequence is that governability is a dynamic property: a system recoverable today may no longer be tomorrow, by simple accumulation of layers and automatisms, without any explicit decision ever having removed the circuit breaker. Which no point-in-time audit will see, because it certifies a state and the problem is a trajectory.

The clinical terrain offers a test of this grid. As long as this grounding is not established, the example remains a working hypothesis, not a proof.

## 7. What to ask, what can be measured, and where the thesis stops

The regulatory consequence is simple to state and costly to hold. The existing texts (GDPR Article 22 on automated decision-making, AI Act Article 14 on effective human oversight, normative frameworks such as the NIST AI RMF or ISO/IEC 42001) are already evolving toward notions of supervision effectiveness. The critique therefore does not target the texts, which would be a straw man, but their operationalization: as long as an audit verifies the presence of a human rather than the recovery capacity of the system, it certifies a stage set.

It remains to explain why presence so regularly prevails over recovery. It is not merely a flaw of regulatory design; it is an asymmetry of verifiability. The presence of a human is observable at a glance, auditable on the record, legally opposable, and almost free to

ascertain: a signature, a timestamp, a log suffice. Recovery capacity, by contrast, is not ascertained, it is tested: it must be simulated, exercised, its restoration and its timing measured. The regulator, like the organization, spontaneously optimizes what is verified at low cost. The signature wins because it is *cheap to verify*, not because it is effective. Any proposal that ignores this verification cost will remain a wish.

One must also anticipate the perverse effect. If a regulator adopted recovery capacity as an indicator, organizations would optimize the indicator without the thing: circuit breakers that exist on paper, rollback procedures never tested. This is Goodhart's law. The riposte does not lie in a better presence indicator, but in a test of effect: recovery is not declared, it is put to the test, as one tests a backup by restoring it and not by verifying that it exists.

This authorizes a beginning of a metric, on condition of treating it as a parameter of an exercised test and not as a declarative checkbox. Four magnitudes are opposable, and all have meaning only when measured on a real exercise:

1. the observed recovery time between detection and the return into the acceptable space,
2. the fraction of states effectively restored during exercises, relative to the states targeted,
3. the margin before the irreversible threshold, that is, the delay between the moment recovery becomes possible and the moment it becomes vain,
4. the coverage of restoration scenarios actually played, and not listed.

These figures are worth nothing isolated from the protocol that produces them. We thus return to the framing that any datum demands: what they prove (a catch-up machine functioned, here, within this time), what they do not prove (that it will function in the face of an unrepresented perturbation, §4), and why we mobilize them nonetheless (for want of better, an exercised recovery is the only available index of a real recovery). It is the same exigency as the master thesis, applied to measurement: performance under perturbation, not inventory.

Three limits, finally, to be named rather than concealed.

1. The first has been said: one does not demonstrate the capacity to manage a specific unknown perturbation, only the presence of generic recovery capacities put to the test. The uncertainty about the unprecedented is irreducible (§4).
2. The second: as soon as one passes from the individual supervisor to the collective (credit committee, multidisciplinary team meeting, board of directors, this or that advisory committee…), the dominant mechanism ceases to be architectural and becomes political: deliberation, coalition, diffusion of responsibility. That is another object.

3. The third: designing a system that remains recoverable when the unexpected arrives is a question so vast that it calls for separate treatment; this note has only mobilized it as a decisive regime, grazing the heart of the subject.

We sum all this up in a sentence. One does not govern what one watches. One governs what one can recover, and only as long as one has put it to the test.

## 8. Conclusion

The fundamental question is not whether a system is watched, nor even whether it is resilient.

It is whether an authority effectively retains the capacity to alter its trajectory when the assumptions that presided over its design cease to be true.

Governability is not the existence of control mechanisms. It is embodied in the persistence of a power of action under perturbation.

A system ceases to be governable not when it commits an error, but when no realistic intervention can any longer inflect its evolution.

It is only at that instant that supervision becomes a stage set: an observable presence but without demonstrated capacity to modify the system's behavior. Control subsists as an organizational, legal or regulatory attribute; it has ceased to exist as effective power of orientation. The authority remains formally present, but its action no longer alters the real trajectory of the system.