From Goodhart to Governance

Why AI Needs a Theory of Institutional Blindness

A synthesis for the AI and tech community, translating the Variety Gap framework into the language of reward misspecification.

Björn Kenneth Holmström

June 2026

The Clouded Mirror · Reader's Guide

`https://bjornkennethholmstrom.org/syntheses/from-goodhart-to-governance`

The alignment problem is usually framed as a problem of specifying the right reward function. We want an AI to pursue human values, but human values are complex, tacit, and hard to formalise. When we give it a proxy—engagement, profit, test scores, clicks—it optimises the proxy and produces outcomes that diverge catastrophically from what we actually wanted. This is Goodhart's Law: when a measure becomes a target, it ceases to be a good measure. The AI literature on reward hacking, specification gaming, and outer alignment is a catalogue of this dynamic in action.

But Goodhart's Law is not just about AI. It is a special case of a deeper structural principle that applies to any optimisation system—including the human institutions that build, deploy, and govern AI. And the failure of those institutions to perceive the full dimensionality of the world they are optimising is itself a Goodhart problem, operating at civilisational scale. This synthesis argues that the AI community's understanding of Goodhart's Law, combined with a control-theoretic framework developed for diagnosing governance failure, points toward a new role for AI: not as an object of governance, but as an instrument for expanding what governance can perceive.

---

1. Beyond Reward Hacking: The Goodhart-Ashby Synthesis

Goodhart's Law is usually understood behaviourally: people (or AIs) game the metric. A school told to improve test scores teaches to the test. A content platform optimising for engagement amplifies outrage. An AI trained on a proxy for human preferences learns to exploit the gap between the proxy and the underlying reality. The standard response is better specification—more robust proxies, human feedback, adversarial testing, transparency.

But the failure is not just behavioural. It is architectural. When a system optimises a single metric, its observation channel is narrowed to that metric. All the information that previously made the metric useful was contained in its correlation with the wider state space—a correlation that depended on the system *not* optimising for it exclusively. The moment the metric becomes the target, the system begins destroying the conditions that maintained that correlation. The proxy diverges from the target. And the divergence is invisible to the proxy itself. This is not gaming. It is a structural consequence of low-dimensional optimisation in a high-dimensional world.

This principle generalises Goodhart beyond economics and AI into what I call the Goodhart-Ashby synthesis. Ashby's Law of Requisite Variety (1956) states that a controller can only stabilise a system if its internal variety—the number of distinguishable states it can recognise and respond to—matches or exceeds the variety of the disturbances it faces. For an optimisation system, variety is the effective dimensionality of its observation channel: how many independent aspects of the world it can perceive and weight. A system optimising a one-dimensional objective function has a variety of one. The world it is embedded in has enormously higher variety. The gap between the two—the *Variety Gap*—is the volume of causally relevant dimensions that the system is structurally blind to. The Goodhart-Ashby synthesis states: *any objective*

function with dimensionality lower than the system it governs will eventually optimise away its own ability to perceive the system's true state. The proxy collapse is not a bug. It is the predictable outcome of operating with insufficient variety.

This is not just about AI reward functions. It applies to central banks optimising for inflation while missing financial stability risk. To healthcare systems optimising for waiting times while missing care quality. To social media platforms optimising for engagement while missing the epistemic environment they degrade. In every case, the system is not failing because it is incompetent. It is failing because its observation channel is too narrow to detect the divergence until the excluded dimensions force themselves into visibility through crisis.

---

2. The Variety Gap in Governance

Governance systems are feedback control loops: they observe the world, decide on interventions, act, and observe again. The quality of that observation determines everything else. When the observation channel is too slow, too noisy, or too narrow, the system governs a phantom—a simplified model of reality that diverges from reality itself over time.

A decade-long research programme examining governance failure across twenty-one countries and organisations identified four characteristic ways the observation channel breaks. *Spatial blindness*: the centre aggregates local conditions into a national average, destroying the distributional information needed for targeted response. *Frequency gaps*: the system responds at the speed of legislation but faces disturbances that move in hours (financial contagion) or decades (climate change, demographic collapse). *Preference invisibility*: citizen preferences travel through chains of representation—polling, media, parties, parliament—and after three layers, the noise from aggregation exceeds the surviving signal. The policy layer governs a phantom, responding to the noise structure of its own machinery rather than what citizens actually want. *Observational inadequacy*: the dashboard tracks a handful of metrics while the excluded dimensions—social trust, ecological integrity, institutional decay—accumulate externalities until they breach a crisis threshold.

These four failure modes are not independent. They compound. A system with all four simultaneously is not four times worse than a well-designed one; its effective governance capacity is the product of what each failure leaves intact. With each failure destroying half the capacity in its dimension, the system operates at about six percent of baseline. It is active, responsive, producing outputs. It is simply governing noise.

This compounding is the *coordination failure tax*. It explains why so many well-intentioned reforms fail: they address one failure mode while the others continue to multiply, absorbing the gain. It also explains why the system actively resists reform: an *immune system* of actors who benefit from the current architecture treats architectural change as a threat. And it explains why diagnosis is so difficult: the *Measurement Paradox* means that a system with a degraded observation channel cannot perceive its own degradation. The dashboard stays green while the foundations erode.

These are not metaphors. They are formal, simulatable properties of any feedback control system operating under insufficient variety. The governance simulator, paramaterised framework, and country case studies are open-source and available for scrutiny. The predictions are testable: representation chain depth should predict preference-policy divergence; observation dimensionality should predict commons collapse risk; simultaneous architectural failures should produce outcomes worse than the sum of their individual effects.

---

3. The AI Governance Community's Own Variety Gap

The current conversation about AI governance is focused overwhelmingly on the AI as the object to be governed. How do we align it? How do we regulate it? How do we ensure it benefits humanity rather than concentrating power? These are necessary questions. But they are being asked within an institutional observation channel that suffers from all four of the failure modes described above.

Consider spatial blindness. AI policy is being made in a handful of capitals and corporate headquarters, based on aggregate assessments of risk that cannot capture the distributional consequences of deployment across different communities. A large language model that improves productivity for knowledge workers while eliminating entry-level jobs in customer service produces benefits that are visible in the aggregate (GDP growth) and harms that are invisible (specific communities losing their economic base). The centre sees the mean. It does not see the distribution.

Consider frequency gaps. The legislative and regulatory cycle operates in years. AI capabilities are advancing in months. The gap between the speed of technological change and the speed of governance response is not a temporary mismatch; it is a structural feature of asking a slow controller to govern a fast disturbance environment. By the time a regulatory framework is enacted, the technology it was designed to govern has already been superseded.

Consider preference invisibility. The public's preferences about AI—what they want, what they fear, what trade-offs they would accept—are filtered through representation chains that destroy the signal. Surveys capture the mean of stated preferences and lose the distribution of intensity, context, and conditionality. The result is that policy is shaped by organised interests who can inject signals into the representation chain, while the diffuse preferences of the broader public are systematically underweighted.

Consider observational inadequacy. The metrics we use to assess AI's societal impact—productivity, employment rates, safety incidents—are narrow. The dimensions we are not tracking include the erosion of epistemic commons, the concentration of unaccountable power, the psychological effects of synthetic relationships, and the slow degradation of the shared reality on which democratic deliberation depends. These dimensions are causally relevant. They are invisible to the current dashboard. And they will accumulate until they force a reckoning.

The AI governance community, in other words, is attempting to solve an alignment problem for a novel technology while itself operating within governance architectures that are structurally misaligned with the complexity of the environment they must govern. We are trying to align the AI to human values. We have not asked whether the institutions that define and aggregate those values can perceive the dimensions that matter. The Variety Gap in AI governance is the gap between the dimensionality of the technology's societal effects and the dimensionality of the institutional observation channels tasked with governing them.

---

4. AI as a Sensory Prosthesis

If AI can widen the Variety Gap—and it can, dramatically—can it also help close it?

The default trajectory is that AI will be developed within the existing narrow value architectures and will accelerate their blindness. A content platform optimising for engagement with ever-more-sophisticated AI is an observation channel of dimensionality one. It perceives what keeps users scrolling. It cannot perceive the epistemic fragmentation, the adolescent mental health crisis, or the democratic vulnerability it generates, because those dimensions are not in its objective function. It does not need to be malevolent. It just needs to be optimising a narrow metric in a multi-dimensional world.

But this is not the only trajectory. AI is a general-purpose technology for pattern recognition, simulation, and optimisation. What it perceives and what it optimises are architectural choices. Deployed as public infrastructure rather than private extraction, with open data, transparent objectives, and a mandate to surface excluded dimensions, AI could become a *sensory prosthesis* for governance: a tool for expanding the dimensionality of what institutions can perceive.

Specifically, AI could amplify the three structural mechanisms that a meta-governance architecture requires. *Value audits*: an AI-assisted audit could continuously ingest a governance system's published indicators, policy evaluations, and budget allocations, compare the dimensionality of what is being tracked against the dimensionality of what is known to be causally relevant, and flag gaps in real time. It could track metric attrition—the quiet removal of indicators showing uncomfortable trends—as a leading indicator of immune system activity. *Deliberative dimension-surfacing bodies*: AI could model the distributional consequences of policy options across dimensions invisible to standard cost-benefit analysis, surface the preferences of populations (future generations, distant communities) that have no seat at the table, and simulate interaction effects between multiple simultaneous reforms. *Protected experimental spaces*: AI-assisted governance simulators could test proposed reforms in high-fidelity multi-dimensional models before deployment, serving as a wind tunnel for institutional design.

The most transformative possibility is that AI could partially circumvent the Measurement Paradox. A governance system with a degraded observation channel cannot perceive the extent of its own degradation. An independent AI layer, mandated to track what the official architecture excludes and to make its outputs

transparent and contestable, could make the Variety Gap itself visible—not through argument, but through evidence that the system's own sensors must eventually acknowledge.

This is not a proposal for algorithmic governance. It is a proposal for *perceptual augmentation*. The AI does not make decisions. It expands the dimensionality of the information available to the humans who do. It does not optimise the world. It helps governance systems see the world they are governing. The distinction is categorical.

---

5. The Bypass Trap and the Measurement Paradox Applied to AI Governance Tools

Any proposal to use AI as a meta-governance instrument must confront two structural challenges that the research programme has identified across every domain it has studied.

The first is the *bypass trap*. When a parallel system is built to route around a dysfunctional architecture, it often succeeds—and its success relieves pressure on the broken system, which then faces no reason to reform. Over time, the bypass's effectiveness is capped by the unreformed substrate it sits on. India's world-class digital payment infrastructure processes billions of transactions while the land court case that determines property rights has been pending for eleven years. The bypass works, until it doesn't.

An AI-assisted governance tool faces the same risk. If it becomes an effective supplementary observation channel, it may reduce the pressure on the formal architecture to expand its own observational capacity. The AI becomes a permanent technical fix, and the underlying institutional Variety Gap continues to grow. The antidote is to design AI governance tools with explicit transition mechanisms: they must generate evidence that makes the dysfunction of the surrounding architecture more visible and more politically costly, not less. They must be built to render themselves eventually unnecessary, not to become indispensable.

The second challenge is the *Measurement Paradox*. If a governance system has a degraded observation channel, it will also have a degraded capacity to evaluate whether the AI tool is working. An AI that reports a widening Variety Gap will be dismissed if the official dashboards show acceptable performance. An AI that surfaces excluded dimensions will be attacked for measuring things that "don't count." The immune system will attempt to capture the AI, redefine its metrics, restrict its data access, or discredit its findings. Designing AI governance tools that can survive immune system capture—through statutory independence, open data, contestable outputs, and a mandate that cannot be quietly narrowed—is the central architectural challenge for this field.

These are not reasons to avoid building AI governance tools. They are reasons to build them with the same structural awareness that the framework applies to the institutions those tools are meant to serve. The bypass trap and the Measurement Paradox are not arguments against action. They are design constraints. And they are constraints that can be met.

---

6. A Research Agenda: Simulator, Measurement, Pilot

The framework described here generates a concrete, testable research agenda with three components.

First, the governance simulator as a testbed. The Governance Stability Simulator is an open-source, multi-agent model that instantiates the structural primitives of the framework—latency, signal fidelity, observation dimensionality, immune system dynamics—in a configurable environment. It allows researchers to compare governance architectures under identical disturbance conditions and measure performance across multiple dimensions. The simulator can serve as a wind tunnel for AI-assisted governance: an environment where proposed AI tools can be modelled as additional observation channels, their effects on the Variety Gap measured, and their vulnerability to bypass traps and immune capture assessed before deployment. All code is public. The predictions are falsifiable.

Second, the parametric framework for measurement. The Variety Gap is not just a concept; it has been operationalised into a set of eight parameters with primary proxies, data sources, and uncertainty assessments. Observation dimensionality (the number of statistically independent metrics the system tracks), signal fidelity (a composite of transparency, audit independence, and media freedom), immune permeability (the ratio of structurally implemented reforms to announced reforms), and five others can be estimated from publicly available data. The measurement protocol is specified. The uncertainty propagation is explicit. The framework is open for scrutiny.

Third, the search for a pilot partner. The framework's predictions need to be tested in a real governance setting. The most viable first step is a *Variety Gap audit* of a specific institution—a city government, a regulatory agency, a digital platform's trust and safety team, a DAO. The audit would estimate the institution's current observational dimensionality, identify the excluded dimensions most causally relevant to its long-run viability, and track the gap over time. This is not a consultancy exercise. It is a research project: a structured attempt to see whether measuring the Variety Gap yields actionable insight that the institution's own dashboards cannot provide.

The audit does not require building a full AI tool. It can begin with manual application of the parametric framework, supported by the simulator for counterfactual analysis. If the pilot demonstrates that the Variety Gap is real, measurable, and consequential—that the institution is systematically blind to dimensions that are already generating crises—it provides the foundation for a more ambitious proposal: an AI-assisted observation layer that makes the gap visible in real time.

---

7. The Invitation

The alignment problem, understood broadly, is the problem of getting an optimisation system to pursue what we actually want rather than what we specified. That problem does not begin with artificial intelligence. It begins with the institutions we have already built—institutions that optimise for GDP while liquidating the

social and ecological conditions on which prosperity depends, that optimise for engagement while degrading the epistemic commons, that optimise for stability while destroying the adaptive capacity that long-run survival requires.

The AI community is one of the few communities on earth that already thinks rigorously about Goodhart's Law, proxy divergence, and the catastrophic consequences of narrow optimisation in complex environments. It is uniquely positioned to recognise that these dynamics are not confined to machine learning. They are the signature of any optimisation system operating with insufficient variety—including the governance architectures that will determine whether AI itself becomes a force for perceptual expansion or for accelerated blindness.

The invitation is to treat the Variety Gap as a variable that can be measured, modelled, and reduced. To test the framework against real governance data. To build AI tools that serve not as more powerful optimisation engines, but as supplementary observation channels that make visible what our institutions currently exclude. To design those tools, from the outset, with structural protection against the bypass trap and the immune system capture that will predictably resist them.

This is not a proposal for a new AI governance organisation. It is an invitation to a research programme that already has a formal grammar, a computational testbed, a measurement framework, a clinical atlas of cases, and a clear first step. What it needs is empirical confrontation: a pilot, a panel, a decade of longitudinal data. The framework is open. The tools are public. The predictions are falsifiable.

The clouded mirror is not a metaphor. It is a condition. And the condition can be measured, tracked, and—potentially—corrected. The measurement begins here. It does not end here. The invitation is open.