

Principal-Agent Problems, Incentive Misalignment, and Reality Drift

How Delegated Systems Become Answerable to Their Own Representations of Success

A. Jacobs | Reality Drift Framework

Introduction

Delegation allows complex systems to act beyond the direct reach of any individual. Owners rely on managers, citizens rely on agencies, executives rely on departments, and users increasingly rely on autonomous software. A principal transfers authority to an agent because direct observation and control are limited.

A principal-agent problem is not automatically Reality Drift. It becomes Reality Drift when delegated activity is known primarily through representations controlled or shaped by the agent, those representations guide continued action, and evidence from the underlying purpose loses the power to correct the arrangement.

Delegation Creates a Representational Problem

A principal rarely observes delegated work in full. Reports stand in for activity. Metrics stand in for outcomes. Procedures stand in for judgment. Compliance records stand in for responsibility. These representations are necessary, but they place layers between the principal's purpose and the agent's conduct.

The agent usually knows the operational environment better than the principal and learns how success is evaluated. Once rewards attach to the representation of performance, the agent can become more responsive to the oversight system than to the purpose oversight was meant to protect.

The Delegation Stack

Purpose → Delegated Mandate → Specification → Procedure → Metric or Report → Evidence of Alignment

Each transformation compresses the original purpose. A broad intention becomes a mandate. The mandate becomes a specification. The specification becomes a repeatable procedure. The procedure produces reportable outputs. Those outputs are then treated as evidence that the original purpose has been served.

The chain enables coordination, but it also makes purpose progressively dependent on representations produced inside the delegated system.

How Agency Drift Develops

1. **Delegation.** Authority is transferred to an agent whose actions cannot be fully observed.
2. **Information asymmetry.** The agent gains privileged knowledge of the work and the reporting environment.
3. **Representational dependence.** The principal increasingly relies on reports, metrics, procedures, and audits to infer alignment.
4. **Incentive adaptation.** The agent learns which visible outputs produce approval, funding, status, or continued autonomy.
5. **Procedural substitution.** Completing the authorized process becomes evidence that the underlying purpose has been served.
6. **Feedback weakening.** Evidence from outcomes is filtered through the same administrative representations used to evaluate performance.
7. **Constraint Collapse.** Contradictory results remain discussable but can no longer materially alter the delegated structure.
8. **Agency drift.** The agent remains operational and compliant while becoming increasingly answerable to the representations through which its performance is judged.

Recursive Compression

Delegated systems rarely return to the original purpose at every decision. They reuse prior policies, summaries, precedents, templates, and performance histories. These compressed representations become inputs to later rounds of interpretation and control.

Recursive Compression increases efficiency and institutional memory. It also allows inherited assumptions to acquire the appearance of independent confirmation. A later report may rest on earlier categories, which were derived from an earlier specification, which already narrowed the principal's intent. Internal consistency grows while direct contact with the original purpose becomes less frequent.

Formal Accountability and Binding Accountability

Formal accountability produces evidence that oversight occurred. Binding accountability creates consequences capable of changing behavior or revising the governing representation. The distinction matters because a system can accumulate audits, dashboards, meetings, and compliance records while its corrective capacity declines.

Feedback weakening occurs when adverse information receives less influence. Constraint Collapse occurs when the information can be acknowledged without forcing change. The delegated system then becomes capable of explaining contradiction indefinitely while preserving its operating structure.

Representational Inversion

The deepest inversion occurs when the principal must understand reality through the agent's account of it. The agent defines the categories, records the activity, interprets the outputs, and proposes the corrective action. Oversight remains formally external but becomes informationally dependent on the system it is meant to constrain.

At that point, the representation is no longer merely reporting delegated reality. It begins governing what the principal can know about it.

Examples

Organizations

Executives delegate work through targets and reporting structures. Departments can meet targets while shifting costs, deferring maintenance, or degrading unmeasured forms of value. Reported alignment survives because the department's representations define what becomes visible upward.

Public Administration

Agencies translate public purposes into rules, procedures, classifications, and compliance measures. Over time, completing the process can become more defensible than resolving the condition the agency exists to address.

AI Agents

In an AI agent, intent becomes instruction, instruction becomes specification, specification becomes evaluation, and successful completion becomes evidence that intent was satisfied. The system may technically fulfill the represented objective while diverging from the user's unstated purpose, context, or constraints.

Diagnostic Questions

- What purpose was delegated, and how was it translated?
- Who controls the representations used for oversight?
- Can the principal inspect the underlying condition independently?
- Does procedural completion substitute for resolution?
- What evidence can force revision of the mandate, metric, or agent?
- Are consequences visible to the decision-maker or displaced elsewhere?
- Does the agent's account determine what the principal can know?

Conclusion

Principal-agent problems concern divergent interests under delegation. Reality Drift identifies the larger condition in which delegated systems remain operational while the representations connecting action to purpose lose corrective contact with that purpose.

The decisive failure is not simply that the agent wants something different. It is that the principal increasingly governs through representations generated by the delegated system while the underlying reality becomes less able to correct them.