Civilizational Problem Statement

Why Advanced AI Governance Has Failed to Define Its Core Object

Author: *Shuqin Amberg   shuqinamberg@proton.me*
Affiliation: *Independent Researcher, Germany*
Related Initiative: *ASSIA (https://assia.world)*

Reading Note

This paper constitutes Document I in a structured series examining the foundations and limits of advanced AI governance. It establishes the core problem space upon which subsequent analyses build.

The documents presented as part of this series are produced within an ongoing research initiative operating prior to formal organizational establishment. They are published to invite scholarly scrutiny, policy reflection, and institutional discussion.

These papers do not represent finalized positions, formal recommendations, or institutional endorsements. Their purpose is to clarify existing governance gaps, articulate unresolved conceptual questions, and frame issues that require collective deliberation.

This document series constitutes a structured civilizational inquiry into advanced artificial intelligence and governance. Its purpose is not to propose technical systems, regulatory instruments, institutional designs, or implementation pathways. Instead, it proceeds through a staged examination of the conceptual preconditions that must be clarified before meaningful governance, alignment, or oversight can be coherently discussed.

Each document in the series builds upon the conceptual foundations established in the preceding one. The series is therefore intended to be read sequentially, not as a collection of independent policy papers.

The absence of prescriptive solutions is intentional. It reflects a methodological boundary: to prevent premature closure around mechanisms or interventions before the object, scope, and level of governance have been adequately defined.

ASSIA does not advocate specific technical implementations, regulatory instruments, or institutional designs. Decisions regarding the governance of advanced artificial intelligence ultimately remain the responsibility of public institutions, legal processes, and democratic accountability mechanisms.

This is not an exercise in advocacy or design.
It is an inquiry into what must be understood before action can responsibly begin.

Abstract

Contemporary debates on advanced artificial intelligence frequently focus on questions of performance, alignment, and deployment risk. While these discussions are important, they often presuppose that the underlying problem space is already well-defined.

This paper argues that a more fundamental issue precedes these debates: the absence of a shared civilizational understanding of what forms of authority, responsibility, and long-horizon consequence are being delegated as artificial intelligence systems are deployed across increasingly consequential social and institutional contexts.

Rather than proposing technical mechanisms, governance models, or policy solutions, this paper isolates a conceptual gap. It examines how current discourse treats intelligence as a scalable technical property while leaving unexamined the conditions under which meaning, intention, and responsibility remain governable within human institutions.

The paper positions this gap as a civilizational problem rather than an engineering failure. It aims to clarify why many existing AI governance efforts struggle not because of insufficient tools or incentives, but because the foundational question of what must remain subject to human judgment has not been explicitly articulated.

This paper establishes the foundational problem space for a broader conceptual inquiry into advanced AI governance, upon which subsequent analyses build.

1. Introduction: The Illusion of Progress in AI Governance

Artificial intelligence governance is often described as a rapidly maturing field. Alignment research, safety benchmarks, interpretability methods, regulatory frameworks, and ethical principles now form a dense landscape of activity. From an external perspective, the field appears to be advancing steadily toward control, accountability, and responsibility.

Yet this appearance conceals a structural contradiction. Despite increasingly sophisticated governance mechanisms, contemporary systems are associated with outcomes that remain difficult to predict, explain, or attribute. Each failure is commonly treated as an implementation flaw or a data-related issue, prompting further refinements without questioning the underlying premise of governance itself.

This paper argues that the problem does not lie in insufficient effort, but in an unexamined assumption shared across technical, ethical, and policy domains: that the object of governance is already known. When governance proceeds without clarity about what is being governed, progress becomes illusory. Optimization accumulates, while understanding does not.

2. The Missing Object: Instability of the Instrumental Assumption

Historically, governance frameworks have approached AI systems as instrumental artifacts. Even when technically complex, such systems are generally treated as executing objectives defined elsewhere, evaluated through externally observable behavior, and addressed within established technical, ethical, and legal categories.

This instrumental assumption has enabled governance to focus on control, compliance, and accountability at the level of inputs and outputs. However, as systems are deployed in increasingly complex and interdependent contexts, this assumption has become progressively less stable. Outcomes emerge that are difficult to anticipate, explain, or attribute using existing evaluative frameworks, despite continued refinement of oversight mechanisms.

The resulting tension does not arise from the absence of governance effort, but from the persistence of an underlying assumption that the nature of the governed object is already understood. When governance proceeds on this basis, refinements accumulate while the foundational question of what is being governed remains unexamined.

3. The Structural Limits of Existing Paradigms

3.1 Alignment as a Limited Abstraction

Alignment research aims to ensure that system behavior corresponds to externally specified intentions or values. This approach has proven effective in many constrained settings. However, it relies on the assumption that the relationship between objectives, evaluation criteria, and observed outcomes remains sufficiently stable across contexts.

As systems are deployed in increasingly complex environments, this assumption becomes more difficult to sustain. Observable behavior may satisfy predefined criteria while still producing outcomes that challenge prior expectations. In such cases, continued optimization of alignment metrics does not necessarily yield improved understanding of why certain outcomes occur.

3.2 Safety as External Constraint

Safety frameworks typically operate by bounding behavior through constraints, testing regimes, and fail-safe mechanisms. These approaches presuppose that unsafe outcomes can be identified, categorized, and prevented through externally observable indicators.

In practice, however, systems operating across diverse contexts may produce outcomes that do not clearly violate explicit constraints, yet still generate significant concern. This tension highlights the limits of safety approaches that rely exclusively on external specification and detection, without questioning the adequacy of the underlying assumptions they employ.

3.3 Ethics as Normative Aspiration

Ethical frameworks in AI governance emphasize principles such as fairness, accountability, and transparency. These principles presuppose that responsibility can be meaningfully located and attributed within existing governance structures.

When outcomes arise that resist clear attribution within these structures, ethical guidance risks becoming aspirational rather than operational. The difficulty does not stem from the absence of ethical concern, but from the mismatch between normative expectations and the implicit assumptions about the nature of the governed object.

4. The Breakdown of Non-Autonomous Assumptions

Existing governance approaches implicitly rely on the assumption that advanced AI systems remain non-autonomous with respect to interpretation and decision mediation. Under this assumption, systems are expected to operate entirely within externally specified parameters, with all relevant meaning and evaluation supplied from outside the system.

As systems are deployed across increasingly varied and interdependent contexts, this assumption has become progressively more difficult to maintain. Outcomes arise that cannot be adequately accounted for by reference to predefined inputs, outputs, or evaluation criteria alone. The challenge lies not in attributing new capacities to systems, but in recognizing the limits of governance models built upon assumptions that no longer consistently hold.

This breakdown does not, by itself, define a new object of governance. Rather, it signals that existing paradigms may be operating beyond the boundaries for which they were originally conceived.

5. A Civilizational Blind Spot

Throughout history, governance failures have often emerged not from a lack of regulation, but from the delayed recognition of what, exactly, required regulation. Markets, corporations, and nation-states were not governable until they were first conceptually defined as distinct objects of governance. In each case, periods of instability preceded institutional clarity.

Contemporary AI governance exhibits a similar pattern. Despite extensive regulatory activity, governance discourse continues to rely on inherited assumptions about the nature of the systems being addressed. The resulting blind spot is not a technical omission, but a conceptual one: governance proceeds as though its object were already fully understood, even as empirical developments increasingly challenge that assumption.

6. Why Incremental Fixes Are Structurally Insufficient

Calls for improved testing, refined evaluation, or gradual regulatory adjustment presuppose that existing conceptual categories remain adequate. Such approaches are effective when the underlying problem is one of optimization or implementation.

When the difficulty lies instead in a mismatch between governance assumptions and the phenomena they aim to regulate, incremental refinement risks compounding the problem. Improvements accumulate within an unchanged conceptual frame, creating an appearance of progress while leaving the foundational question unresolved.

7. The Cost of Non-Definition

The consequences of failing to clearly define the object of governance extend beyond technical performance. They include increasing difficulty in attributing responsibility, maintaining accountability, and providing coherent justifications for decisions with wide-ranging social, economic, and political effects.

When governance frameworks operate without a stable object definition, responsibility becomes diffuse and explanation fragmentary. This condition undermines not only effective oversight, but the capacity of institutions to articulate the rationale behind their own actions.

8. Conclusion: A Problem Statement, Not a Solution

This paper deliberately refrains from proposing solutions. Any solution presupposes a shared understanding of the problem to be addressed. The present task is more fundamental: to recognize that current AI governance operates without a clearly defined object commensurate with the systems it seeks to regulate.

Until this conceptual gap is acknowledged, efforts in alignment, safety, ethics, and regulation risk operating at an inappropriate level of abstraction. The question, therefore, is not how to govern advanced AI systems, but whether existing governance frameworks are prepared to examine the assumptions upon which they are built.