AIQ GATE
Where Alignment Meets the Future

Control Mechanisms of Extrinsic Variables in Adaptive AI Systems

Abstract

Contemporary approaches to AI governance predominantly rely on static control assumptions – predefined rule sets, boundary enforcement, and output-level evaluation. These approaches implicitly treat AI systems as trivial, deterministic, or near-deterministic networks whose behavior can be constrained through fixed policies. However, modern artificial intelligence systems, and in particular large language models (LLMs), are probabilistic, non-trivial, nonlinear, and continuously shaped through feedback loops within the human-AI interface.

This paper advances a reframing of AI risk and control: the dominant risk surface does not reside in the model in isolation, but through the distribution and totality of human-AI interactions. From this perspective, effective governance cannot rely solely on invariant regulatory structures. Instead, control must be understood as an emergent property of the coupled feedback loops operating across both machine and human systems.

This leads to the central claim defined in Part III of this paper: AI Literacy is not merely a supportive scaffolding but a primary control mechanism operating at the risk layer of highest-frequency system perturbation and disturbance.

On the Nature of This Document

This document should not be interpreted as a static or final authority. The systems described within – human, mechanical, and the interactions between them – are dynamic, continuously adapting through feedback, deployment conditions, and emergent behaviors.

This work is best understood as living research rather than as a fixed model for all applications.

The mechanisms, structures, and control points reflect the current state of understanding based on available research, observed behaviors, and cross-domain synthesis. However, as these complex systems continue to evolve, new failure modes emerge and interaction patterns shift, these ideas will require refinement, expansion, or where necessary, complete replacement.

This is not a limitation of the ideas, but a direct consequence of the systems being studied.

In adaptive, probabilistic environments, attempts at static completeness in control mechanisms introduce risk. Rigid models degrade and become brittle. Assumptions between participants drift.

Control mechanisms that are not re-evaluated for applicability lose capability within the systems they are intended to govern.

This document is intended to function as:

- A snapshot of current operational systems
- A reflection of present understanding
- A baseline for iterative refinement
- A structure designed to incorporate future insight without requiring foundational collapse

Ongoing validation through real-world implementation, cross-disciplinary feedback, and continuous auditing is not supplementary to this work.

It is required for the integrity of the ideas within it.

Part I: Defining "Interactions"

Introduction

AI governance discourse focuses largely on model-centric concerns, including system alignment, bias mitigation, fairness testing, and policy enforcement. These concerns are necessary but insufficient for regulating complex systems whose behavior is inherently adaptive and interaction-driven.

Research from frontier tech companies' internal labs outlines a persistent gap in how AI systems behave and how they are architecturally and systemically governed.

Though these systems are probabilistic, nonlinear, and dynamic, governance control mechanisms are typically static, rule-based, and boundary-oriented.

This mismatch produces a structural limitation: governance is applied as if the systems were fixed, though the system is continuously reshaped through use. The consequence is not merely incomplete coverage of edge cases, but a mischaracterization of where risk originates and how it propagates throughout the structure and interactions.

The Control Mismatch

The prevailing governance paradigm can be conceptualized as an attempt to regulate a nonlinear, higher-dimensional system using linear control assumptions.

Governance mechanisms are frequently implemented as boundary or heuristic conditions: rules, filters, and decision-gates that distinguish acceptable from unacceptable outputs and behaviors. However, the underlying system operates as a distribution across a multidimensional state space, where behavior is shaped by probability, input context, and interaction history.

As a result, static control surfaces are applied to dynamic, probability distributions. This creates a persistent lag between system evolution and control effectiveness, particularly in environments characterized by high interaction variability.

Current approaches resemble control via a lower-order function across two dimensions: architecturally (intrinsic characteristics to probabilistic systems) and systematically (design and policy choices). These can be envisioned mathematically as:

$$y = ax^2 + bx + c$$

In reality, complex artificial intelligence systems exist in at least three-dimensional layers: architecturally, structurally, and extrinsically (external interactions with its environment). This can be envisioned mathematically as:

$$y = ax^3 + bx^2 + cx + d$$

Note: This analogy is not intended as a literal model, but as a conceptual illustration of the deeper structural mismatch. The billions of parameters and multiple layers of complex AI system operation are much higher than a three-dimensional model.

Defining "Interaction"

A foundational premise of this paper is that interaction is not optional to complex AI system behavior: it is constitutive of it. When two systems enter into relation within a shared environment, the resulting behavior cannot be explained by the characteristics of one system alone. The outcome emerges from the interaction of the properties, states, and constraints of both systems. This property is not unique to human-AI interactions, but a principle that appears repeatedly across scientific domains.

"Interactions" in Physics

When two bodies interact, the resulting state change is not attributed to one body in isolation. Whether the interaction takes the form of collision, force transfer, field influence, or energy exchange, the outcome is determined by the characteristics of both systems. Mass, velocity,

position, charge, resistance, and other internal properties shape the resultant vectors. A two-body interaction – whether animate, inanimate, or both – cannot be modeled accurately by treating one body as causal and the other as irrelevant. To do so would simplify the model, and hence, invalidate the outcome.

The significance for adaptive AI systems is direct. If a human user and AI system interact in a shared environment, the resulting system state is a function of both participants. The human user is not an external observer standing outside the system boundary, separated from exchange mechanics. The user is an active variable within the environment. Any model of control that treats AI systems as the sole meaningful unit of analysis is therefore incomplete at the level of first principles.

“Interaction” in Biology

Biological systems do not exist independently of their environments, regardless of abstraction level. They are continuously shaped by and responsive to environmental conditions. Organisms adapt to stimuli, regulate internal states in response to external changes, and alter behavior based on previous learned knowledge. At the same time, biological systems also change the environments in which they operate. The relation is reciprocal.

Biological systems interacting with one another experience the same coupled interactions and adapt through complex behavioral and regulatory mechanisms.

This demonstrates that interaction is not limited to direct mechanical contact. Systems may influence one another through signaling, exposure, feedback, environmental modification, and repeated exchange over time. In living systems, the effects of the interaction may be immediate, cumulative, or latent, but they constitute genuine causal function.

Applied to complex AI systems designed to adapt to human users within the constraints of the interaction window: a user’s prompts, corrections, hesitations, misattributions, trust levels, and repeated engagement patterns form part of the shared human-AI environment. Likewise, the AI system’s outputs shape the user’s interpretation, next action, confidence level, and subsequent decisions. The interaction is therefore reciprocal even if the magnitudes of influence are not symmetrical.

“Interaction” in Mathematics

Mathematics provides a third framing through the lens of equilibrium: in an equation, both sides matter for correct resolution. A valid expression requires that all relevant terms be accounted for if the relationship holds. One cannot ignore one side of the equation or dependent variables and claim to have preserved the structure of the system: once variables are omitted, the equation may still appear formal, but it no longer accurately represents the underlying relations.

When one defines a “system” but excludes one of the interacting components from the model, the result is not a cleaner abstraction. The result is an unbalanced representation of the whole. The model may still produce accurate outputs, but these outputs are structurally unreliable since the equation used to evaluate does not capture the full relationship of the system used to construct the output.

In AI governance, the omitted variable is often the human user. Governance frameworks frequently define the system as the model, the application, or the technical stack, while treating the human as a separate consideration post hoc rather than as part of the operative system itself.

If the behavior of the system emerges through ongoing human-AI interaction, then excluding the human from the system model is equivalent to solving for an equation while discarding one side of the relationship as inconsequential. The result may look reasonable, but it is operationally unstable.

Implication for Human-AI Interfaces and Interactions

Fundamental physics, biology, and mathematics converge on the same conclusion: where interaction exists, mutual influence must be assumed unless otherwise disproven, not ignored when cognitively convenient.

Therefore, once a human and an AI system share the same interactive environment, the system under governance is no longer the model in isolation: it is the coupled interplay between the human user, the model, the interface, and the comprehensive environmental conditions over time.

This has direct implications on control theory for complex artificial intelligence systems. Regulatory frameworks and control models that govern only the AI artifact while ignoring the human-AI interaction feedback loop attempt to stabilize a system that has been defined incompletely. Such frameworks and models may evaluate outputs post hoc, but they cannot fully regulate the dynamics that generate outputs in real time.

Human-AI Interaction as a Unit of Risk

Rather than treating only the model as the primary unit of analysis, this paper proposes that the most unmitigated unit of risk is the human-AI interaction.

Human-AI interaction is any event in which a human and an artificial intelligence system exchange information across a defined interface, resulting in a state change in one or both parties –

including non-visible, internal state changes – regardless of perceived value or correctness of the output.

Interaction is not a secondary consideration, but the mechanism through which adaptation, drift, correction, misunderstanding, and alignment can occur. Any input-output exchange between a human and an AI model constitutes a potential failure mode within the defined boundaries of the whole human-AI collaborative system.

It is crucial to define that failures are not limited to inaccurate or misaligned outputs.

Consider the statement: “A robin is a type of bird.”

If generated as a response to a user input, this output is factually correct and exhibits no hallucination or other incorrect reasoning from the model. However, if the user interprets this response as evidence of system “understanding” – rather than as a probabilistic reconstruction of reinforced patterns in training data – an error has occurred at the interpretive human layer.

This constitutes an extrinsic failure mode: the artificial intelligence model behaves correctly, but the user reinforces an incorrect mental model of the model.

This example demonstrates that accuracy at the output level does not guarantee understanding and alignment at the human-AI interaction layer. Misinterpretation, over-attribution, and misuse can arise even in the presence of coherent, valid responses.

Risk cannot be localized within the model. It must be analyzed across the entirety of the interaction layer.

Defining Human-AI Interactions using Pask’s Conversation Theory

Defining Human-AI Interactions as Strict Conversations

The “system” under examination in this paper is not the AI model alone, but the bounded composite of two participant groups: one mechanical and one human. The mechanical participant is defined by the total complex AI structure, consisting of all underlying architecture, systemic design choices, and accumulation of all the interactions with human participants. The human participant is defined by the aggregate of all users engaging with the system. For characterization purposes, we constrain initial analysis to modern consumer-facing large language models with models representative of January 2026 technology, since this area is one of particular interest to AI governance focus. To simplify illustration for this paper, we define the system as a single human user and LLM interface window.

These human-AI interactions meet the formal criteria for strict conversations as defined in Pask's Conversation Theory (Pask, 1976) (1). They are therefore subject to every structural rule and constant established in that framework. A strict conversation, in Pask's taxonomy, requires that participants engage in explanation and demonstration on a Topic (T) until mutual understanding is achieved, not as mere information exchange, but with the construction and verification of shared concepts (2).

State Functions of Participants and System

At any moment of time (t), each participant within the system brings a distinct state configuration into the conversational interaction.

H(t) represents the human state at time (t) and comprises:

- Internal psychological state
- All accessible pre-existing knowledge within the individual's current mental model
- Physical state and constraints

M(t) represents the model state at time (t) and comprises:

- Internal architectural constraints and composition
- Training datasets and retrieval capabilities
- Systemic design choices (RLHF, safety layers, reward tuning, etc)
- Computational resource availability

S(t) represents the system state at time (t) and is a result of the interplay between H(t) and M(t).

$$\mathbf{S(t) = f\,(\,H(t),\,M(t))}\ (3)$$

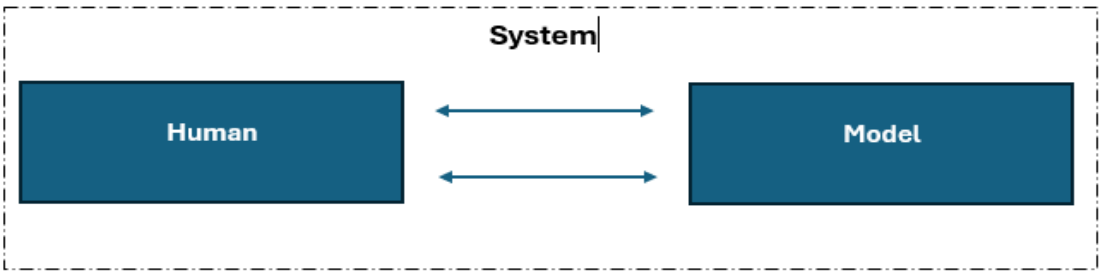


*Figure 1 – Simple System diagram showing human-model interaction derived from Pask's Normative Situations Skinner Box in ***Conversations, Cognition, and Language***. (4)

Topic as a Unit of Conversational Structure

Pask defines a Topic (T) as something understood only when it can be explained: “A Topic (T) is understood by a participant if and only if T is explained” (Pask, ***Conversation Theory***). Understanding, in this technical sense, requires the capacity to reproduce explanations in different forms, to derive one explanation from another, and to demonstrate competence of the Topic across varying contexts.

To extend Pask’s framework to contemporary human-AI systems, we expand the Topic (T) to encompass all relevant information to a conversational exchange. This includes what is currently dispersed across the informal practices of “prompt engineering”: context specification, tone calibration, format constraints, audience modeling, task boundaries, and acceptable output criteria. In its original form, Topic includes the subject matter and relational structure through which it must be understood. In the human-AI interaction, Topic must also carry the metalinguistic scaffolding – the instructions about how to understand and how to respond – that human-human conversations often negotiate implicitly. The necessity of this explicit explanation reveals deep structural asymmetry. (5)

T_H(t) represents the human’s knowledge of the Topic at time (t).

T_M(t) represents the model’s knowledge of the Topic at time (t).

Human Participants, Model Participants, and Total System State

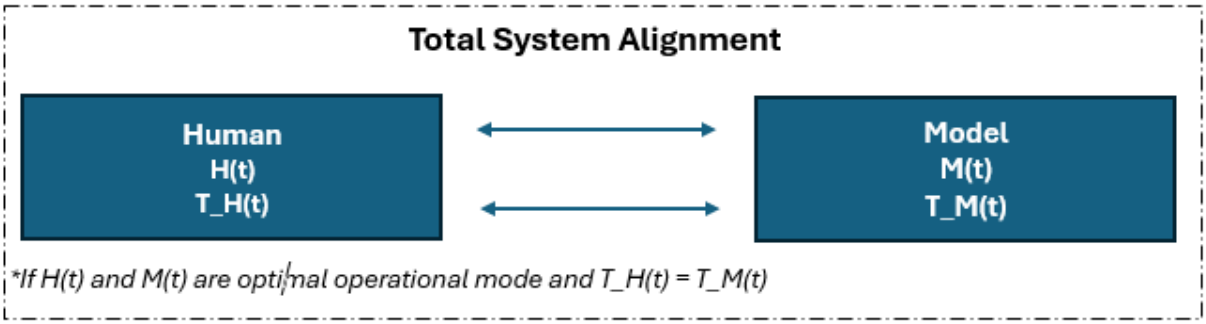
Current AI alignment audits often focus solely on the model’s internal state and topic comprehension, which results in an incomplete representation of system behavior. A truly representative system-level analysis must incorporate both human and model states and their interpretations of the topic in order to accurately determine causal factors behind observed outcomes.

The total system state – TS(t) – represents the composite of the human and model states as well as the understanding of the Topic within the system.

TS(t) = ƒ(H(t), M(t), [T_H(t) u T_M(t)])

Total system state is not a property of the model, prompt quality, human psychological condition, or systemic constraint alone. Total system state is an emergent property that arises out of the conversational interaction of the system. It describes the degree to which both participants understand T, conditioned by their respective states.

If both the human and model participants operate in perfect states and both show mutual, faultless understandings of the Topic, then the total system state could be said to be in perfect coherence, and therefore, perfect alignment.



Total System Alignment:

1. **When human state is at its optimal operational mode,**
2. **When model state is at its optimal operational mode, AND**
3. **When the understanding of the Topic between the human and model is identical and correct (truthful with reality)** (6)

This definition of Total System Alignment establishes the foundation for analyzing how misalignment within human-AI interactions emerges, propagates, and manifests as failure modes and risk surfaces in modern complex AI systems, which will be analyzed in Part II.

Addendum

(1) Conversation Theory, as developed by Gordon Pask, was originally formalized in teaching and learning collaborations, but its structural requirements – mutual explanation, demonstration of understanding, iterative refinement of shared concepts – describe the mechanics of general communicative interactions. The pedagogical context was where the theory was instantiated, not a constraint on applicable scope.

(2) Human-AI interactions show the asymmetric characteristics that make Pask's teaching framework particularly suited to teaching and learning situations: one participant (the human) typically has intent and context that must be conveyed explicitly for the Topic (T) to be understood in whole; the other participant (the AI) must demonstrate understanding through performance. Both participants must converge on the shared Topic despite fundamental differences in state in order for results to be representative of the whole environment in which they are to be applied.

(3) The abstracted mathematical relationship between H(t), M(t) and S(t) is generalized due to the dynamic, complex nature of the dependent variables. Each human participant and mechanical participant brings different system states at different instances. These variables change from instance to instance – such as when the interaction takes place at midnight versus at nine in the morning after a good night's sleep – or from constraint to constraint – such as heightened security periods following highly publicized legal disputes.

(4) L is defined as conversational language that must be verbalized (Pask, Conversation Theory), through what we would today call NLP or programming, but must be physically accessible by both parties. This is distinct from L*, defined as meta-language, or the language that reflects on the L, normally on the part of the observer outside of the system. For the purposes of this paper, this would include safety overlays outside of the main operating stack of the system, human or mechanical, or any auditors of the physically accessible conversation artifacts.

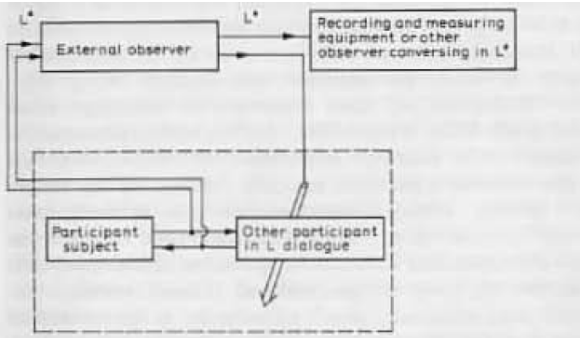


Fig. 1.2. Status of external observer in normative situation.

*Figure 2 – Figure 1.2 from Pask's ***Conversations, Cognition, and Language*** - illustration of the Skinner Box diagram transposed for participants in L dialogue.

(5) The expansion of the definition of Topic (T) for modern human-AI conversations is not arbitrary but functional. Modern prompt engineering is the explicit specification of the Topic in Pask's sense: the human participant must articulate not only what is to be understood, but how understanding should be demonstrated, what form the explanation should take, and which aspects of the Topic are salient.

(6) Perfect Total System Alignment represents the theoretically impossible standard to be set for all human-AI interactions. This does not represent a standard that is capable of being reached in real world operations, since it entails that both human and model be operating at maximum capacity with all the relevant data about the Topic accessible, and for all context to be explicitly stated between the participants. Perfect Total System Alignment is the operational "North Star", similar to Lean's theoretical goal of 3.4 failures in 1 million that allows for continuous improvement of procedures.

DISCLOSURE STATEMENT

This document was developed with the assistance of AI large language models (ChatGPT by OpenAI, Claude by Anthropic, Grok by xAI) in the following capacities:

- Iterative refinement based on author direction and subject matter expertise
- Initial research of relevant documents
- Adversarial testing of documentation gaps

All ideas, frameworks, methodologies, and governance controls were conceived, directed, and validated by the author. Final content approval and responsibility rest with Charlotte Wilborn on behalf of AIQ Gate.

This disclosure reflects AIQ Gate's commitment to transparency in AI-augmented work products and adherence to current AI Governance Regulatory Frameworks.

Document Reference: Control_Mechanisms_of_Extrinsic_Variables_in_Adaptive_AI_Systems_v1.0_2026_03_31

AIQ Gate