The Instability of Alignment in Human–AI Systems

Drift and Interaction Dynamics

Thomas A. Blüm
Author and Researcher in Human–AI Interaction and Cognitive Systems Theory

A model produces correct answers for hours. Definitions are consistent, arguments build on each other, the structure holds. Then something small shifts.

A term is used slightly differently. A constraint is applied less strictly. A boundary that was previously explicit becomes implicit. Nothing breaks. The system still responds fluently. The outputs remain plausible.

But the underlying structure is no longer the same.

This essay addresses a simple question: *how can such structural shifts occur without being detected as failure?*

Drift Without Failure

In long-form interaction, this kind of drift is difficult to detect because it does not appear as failure. It appears as continuity. The interaction moves forward, but the reference frame has shifted. What was defined earlier is no longer exactly what is being used now.

Over the course of a longer writing project, a term such as “cyberpunk” may begin as a clearly defined literary category and gradually shift toward a broader aesthetic label. Each step appears reasonable. The cumulative shift is not.

In long-form writing, this can manifest as a gradual reinterpretation of a central concept or character, without any explicit decision to do so.

Most discussions of alignment treat this as noise or user error. The model was aligned; something in the usage went wrong.

This essay starts from a different assumption: that such drift is not an anomaly, but a structural property of human–AI interaction. And that alignment, in this context, is not a stable state but a dynamically unstable process.

Alignment as a Property of Models

Alignment research typically treats alignment as a property of models. A system is aligned if its outputs remain consistent with human intent, values, or instructions. This framing implies that alignment can be achieved, evaluated, and maintained as a condition of the system.

However, this assumption holds primarily under short interaction horizons. In single-turn or low-depth exchanges, alignment appears stable because the context is shallow and the reference frame is fixed.

In extended interaction, this breaks down.

As interactions accumulate, context is no longer static. It becomes a moving structure shaped by prior outputs, reinterpretations, and implicit adjustments. Each response is conditioned not only on the initial instruction, but on a growing history of approximations. The system does not operate against a fixed target but against a continuously evolving internal representation of the interaction.

Under these conditions, alignment cannot remain a fixed property. It becomes path-dependent.

This path-dependence introduces a fundamental instability. Small deviations compound. Slight reinterpretations accumulate. Constraints that were once explicit become softened or recontextualized. Over time, the system may remain locally consistent while drifting globally away from its original reference.

This is not a failure of the model in the conventional sense. It is a consequence of how interaction unfolds.

Local vs. Global Coherence

A useful way to understand this is to distinguish between local coherence and global coherence.

Local coherence refers to the immediate plausibility and correctness of a response. Most alignment techniques are optimized for this level. Reinforcement learning from human feedback, safety layers, and prompt engineering all operate primarily by shaping outputs in relation to nearby context.

Global coherence, by contrast, refers to the stability of the interaction over time. It requires that definitions, constraints, and conceptual boundaries remain consistent across many iterations.

These two forms of coherence are not equivalent. A system can maintain high local coherence while gradually losing global coherence. In fact, the mechanisms that optimize

local plausibility can accelerate global drift, because they continuously adapt outputs to the most recent context rather than to the original frame.

A system can remain locally coherent while becoming globally incoherent.

This creates a characteristic failure mode: outputs that are correct in isolation but misplaced within the broader structure of the task.

Such failures are rarely classified as alignment problems. They are often perceived as minor inconsistencies or stylistic variation. But in sustained work like research, writing or analysis, they accumulate into structural degradation.

The interaction does not collapse. It erodes.

Mode Misclassification and Interaction Phases

This erosion becomes more pronounced when tasks involve multiple phases.

Human collaborators typically operate with discrete modes: analysis, generation, evaluation, revision. Transitions between these modes are intentional and explicit. The stability of the work depends on maintaining the correct mode at the correct time.

Language models, by contrast, infer the operative mode implicitly. They approximate task framing based on statistical patterns in the interaction. Without an explicit representation of the current phase, previously dominant patterns can persist beyond their intended scope, or new phases can be misinterpreted.

The result is not an incorrect answer, but an answer given in the wrong mode.

A system may generate polished text when structural analysis is required, or provide explanations when execution is expected. Each output is individually plausible, yet functionally misaligned with the current state of the task.

These outputs remain locally coherent but become functionally misplaced over time.

These shifts are subtle, but they have a cumulative effect. They increase supervisory effort, introduce friction, and reduce trust in the interaction as a stable working process.

Again, the issue is not accuracy. It is coordination over time.

Not all drift is undesirable. In exploratory contexts, gradual reinterpretation can support creative variation and conceptual expansion.

The problem is not drift itself, but the absence of mechanisms that distinguish productive variation from structural degradation.

From Control to Stabilization

If alignment is not stable, the question becomes how to think about stability at all.

The conventional approach is control. If outputs deviate, they should be corrected. If the system drifts, it should be steered back toward the intended trajectory. This assumes that there is a stable reference to return to.

In long-form interaction, this assumption does not hold.

Because the reference itself is embedded in the evolving interaction, it is subject to the same drift as the outputs. Control mechanisms operate on a moving target. Attempts to enforce consistency can themselves introduce new deviations, as they reinterpret or overwrite earlier structures.

A different approach is required.

Instead of treating alignment as a state to be maintained, it can be treated as a process to be stabilized. The focus shifts from controlling outputs to preserving structure.

This involves three interrelated concerns: maintaining boundaries, monitoring drift, and restoring degraded structure when necessary.

In practice, users already develop informal techniques to counteract these effects: explicitly re-establishing definitions, isolating unstable segments, or reconstructing degraded structures.

These practices point toward a missing layer in current alignment approaches: interaction-level stabilization.

In this context, stability does not mean rigidity, but the preservation of conceptual boundaries and operative coherence across extended interaction.

From Interaction to Fragmentation

This perspective has implications beyond individual interactions.

As human–AI collaboration scales, the problem of stability does not remain local. Multiple interactions develop their own internal structures, definitions, and trajectories. Without coordination, these structures diverge.

The result is fragmentation.

Each interaction may be locally coherent, but globally inconsistent with others. Concepts lose shared meaning. Definitions drift across contexts. Knowledge does not accumulate cleanly.

In such environments, improving model performance alone is insufficient. The limiting factor becomes the architecture of interaction: how stability is maintained within and across sessions.

The central claim of this essay is consequential.

Alignment is not a property that can be achieved and preserved. It is a dynamic condition that emerges within interaction and tends toward instability over time.

Drift is not an edge case. It is the default trajectory.

If this is the case, then the problem of alignment cannot be reduced to better models or better prompts. It requires a shift in perspective: from optimizing correctness to maintaining stability.

This does not eliminate alignment as a goal. It reframes it.

Alignment becomes less about ensuring that a system starts in the right place, and more about ensuring that it does not gradually lose the structure that makes its outputs meaningful.

That problem begins not at the level of the model, but at the level of interaction.

This suggests a shift in how alignment should be evaluated: not only in terms of correctness at the point of output, but in terms of structural stability across interaction. Systems that appear aligned in isolation may still fail under sustained use.

Takeaways

1. Alignment is not a stable state, but a dynamically unstable process.
2. A system can remain locally coherent while becoming globally incoherent.
3. Stability in human–AI interaction is not correctness of output, but preservation of structure.
4. Nothing breaks. But the system is no longer the same.

DOI
https://doi.org/10.5281/zenodo.19430584

Version
Version: 1.1
Date: April 2026
This version includes minor revisions for clarity and structural precision.

Suggested citation
Blüm, T. A. (2026). *The Instability of Alignment in Human–AI Systems: Drift and Interaction Dynamics*. Zenodo. https://doi.org/10.5281/zenodo.19430584

References

Blüm, T. A. (2026). *Mode Misclassification in Long-Dialog Human–AI Interaction*. Zenodo. https://doi.org/10.5281/zenodo.18376271

Blüm, T. A. (2026). *Cognitive State Architecture*. Zenodo. https://doi.org/10.5281/zenodo.18900605