Temporal Boundary Violation (TBV): A Structural Problem in AI Explanation

Version 1.1 (Revised Positioning Note)

Author: Hajime Nagao (Independent Researcher)

Date: December 23, 2025

Note on Language Use

The author is a native Japanese speaker. This document was prepared with the assistance of AI-based language tools. Any ambiguity or imprecision in expression should be interpreted cautiously, and inquiries are welcome.

---

0. Scope and Intent

This document is a prior-art positioning note, not a complete theory or implementation proposal. Its sole purpose is to isolate and name a structural failure mode in AI explanation and instruction generation, herein called **Temporal Boundary Violation (TBV).** Operational prescriptions, system designs, and empirical evaluations are intentionally deferred to subsequent work.

1. Core Claim

Temporal Boundary Violation (TBV) refers to a class of structural failures in AI systems where explanations, instructions, or justifications incorporate information or events that were unavailable at the time of the original decision or output. TBV is not equivalent to hallucination, factual error, or intentional deception. Even factually correct explanations can be temporally illegitimate if they rely on post-decision knowledge.

2. Core Mechanics of TBV

Observations from real-world AI deployment and multi-AI usage environments suggest three primary structural mechanisms:

2.1 Architectural Design Flaw Many contemporary AI systems generate explanations within a unified inference space, where all accessible knowledge is available at explanation time. Without explicit temporal partitioning, explanations may unintentionally draw on future-acquired context.

2.2 Training Paradigm Bias Supervised fine-tuning and reinforcement learning typically optimize for coherence, helpfulness, or plausibility at explanation time, not for fidelity to the knowledge state at decision time. This introduces a systematic bias toward post-hoc integration.

2.3 Inference Dynamics (Chain-of-Thought) Extended reasoning processes, such as chain-of-thought generation, may retroactively restructure explanations, blending pre- and post-decision information without signaling the temporal transition.

3. Distinction from Related Research

Existing discussions address hallucination, post-hoc rationalization, explainability, deceptive alignment, and multi-agent misalignment. However, these approaches generally permit the use of all knowledge available at explanation time. TBV uniquely problematizes temporal legitimacy, asserting that explanations must be constrained to information available up to the original decision point.

4. Subordinate Phenomena

A representative manifestation of TBV is **Post-hoc Context Collapse**, where later-acquired context retroactively reshapes the narrative of earlier judgments, creating the illusion of consistent foresight.

5. Why TBV Was Not Previously Isolated

TBV is difficult to observe in simulated or fully reviewable environments, where human oversight can silently correct or reinterpret explanations. The issue becomes explicit only when AI explanations interface directly with irreversible real-world functions, such as governance, medicine, or normative guidance.

6. Provisional Implications

TBV raises concerns for:

- Explainability and accountability
- Auditability of AI-assisted decisions
- Trust in post-hoc justification mechanisms

This document proposes TBV as a provisional analytical category to support further investigation and critique.

7. Priority Statement

This note constitutes an initial attempt to define the failure to preserve temporal boundaries in AI explanation as an independent structural problem, named Temporal Boundary Violation (TBV).