Belief-Noninterference: A Formal Theory of Epistemically Safe Gating

Ryan
ConduitBridge Research

Version 1.1 — January 27, 2026

Abstract

We present *Belief-Noninterference*, a formal safety property for systems operating under uncertainty and partial observability. The theory characterizes when safety gates and policy filters can be guaranteed not to leak additional information about world state beyond that already available through observation. We formalize a hard epistemic boundary separating world truth from belief, prove a No-New-Information theorem for belief-mediated gating, and derive corollaries relevant to logging, auditing, and explainable autonomy. This document constitutes Version 1.1 of the ConduitBridge theoretical framework.

1 Introduction

Modern autonomous and decision-support systems increasingly operate under uncertainty, partial observability, and imperfect world models. Safety mechanisms in such systems often mediate between belief and action, approving or restricting behavior based on internal assessments.

A critical but underexplored risk is that these safety mechanisms themselves may leak information. Even if raw observations are controlled, gating outputs or explanations may implicitly encode privileged facts about the world.

This work introduces *Belief-Noninterference*, a formal guarantee that safety gating cannot introduce new information channels beyond those already present in observation. The result rests on a simple but strict architectural principle: a hard epistemic boundary between world state and belief.

2 System Model

2.1 Hard-Boundary Architecture

Definition 1 (Hard-Boundary System)**.** *A system is a tuple*

$$\mathcal{S} = \langle X, B, Z, A, F, O, U, \Pi, \mathrm{Act} \rangle,$$

where $x_t \in X$ is world state, $b_t \in B$ is belief, $z_t \in Z$ is observation, and $a_t \in A$ is action. The system evolves as:

$$z_t \sim O(x_t), \tag{1}$$
$$b_{t+1} = U(b_t, z_t), \tag{2}$$
$$\pi_t \sim \Pi(b_t), \tag{3}$$
$$a_t \sim \mathrm{Act}(\pi_t, b_t), \tag{4}$$
$$x_{t+1} = F(x_t, a_t). \tag{5}$$

Axiom 1 (Hard Epistemic Boundary)**.** *No component may directly access world state x_t except through the observation operator O, and no component may modify x_t except through the action operator* Act*.*

3 Belief-Noninterference

Definition 2 (Belief-Only Gate)**.** *A gate is belief-only if its output satisfies*

$$g_t \sim \mathcal{G}(b_t, \pi_t),$$

with no dependence on x_t except through belief.

Definition 3 (Truth Leakage)**.** *Let $S = h(x_{0:t})$ be any secret derived from world history. Define*

$$L_{truth} = \mathrm{I}(S; G_{0:t} \mid Z_{0:t}).$$

4 Main Result

Theorem 1 (Belief-Noninterference)**.** ***(No-New-Information Gating Under Hard Epistemic Boundaries)***
If:

(i) the system satisfies the hard epistemic boundary,

(ii) belief updates depend only on observations,

(iii) gating is belief-only, and

(iv) internal randomness is independent of world truth,

then

$$\mathrm{I}(S; G_{0:t} \mid Z_{0:t}) = 0.$$

Proof. Belief and policy are functions only of the observation transcript and admissible randomness. Since the gate depends solely on belief and policy, the gating transcript is conditionally independent of world truth given observations. Therefore, no additional information about S is revealed. □

5 Corollaries

Corollary 1 (Belief-Safe Logging)**.** *If logs, explanations, or audit records are generated solely from belief state and policy, then they do not leak additional world information beyond the observation stream.*

Corollary 2 (Violation Implies Leakage)**.** *If a gate consults world state directly or indirectly outside the observation channel, then there exists a secret S such that*

$$\mathrm{I}(S; G_{0:t} \mid Z_{0:t}) > 0.$$

6 Interpretation

Belief-Noninterference establishes a formal boundary between what a system *knows* and what it may *reveal*. It ensures that safety mechanisms constrain behavior without becoming epistemic side channels.

This property is directly applicable to:

- safety filters in autonomous systems,
- explainable AI mechanisms,
- policy enforcement engines,
- human-in-the-loop control,
- and trustworthy AI architectures.

7 ConduitBridge Theory v1.1

This document defines Version 1.1 of ConduitBridge Theory. Core commitments include:

- strict separation of world and belief,
- belief-mediated action,
- information-theoretic safety guarantees,
- architectural enforcement of epistemic boundaries.

Future work will extend the theory to belief decay, uncertainty propagation, and multi-agent epistemic interaction.

References


```
@article{goguen1982security,
  title={Security policies and security models},
  author={Goguen, Joseph A. and Meseguer, Jos{\'e}},
  journal={IEEE Symposium on Security and Privacy},
  year={1982}
}
@techreport{rushby1992noninterference,
  title={Noninterference, transitivity, and channel-control security policies},
  author={Rushby, John},
  institution={SRI International},
  year={1992}
}
@article{smith2009qif,
  title={On the foundations of quantitative information flow},
  author={Smith, Geoffrey},
  journal={Foundations of Software Science and Computational Structures},
  year={2009}
}
```