PKF – Decision Interruption Safety Rule

Author: Eun-Young Lee, M.D.
Affiliation: Pain Medicine · Human-Centered AI · Decision Ethics
Keywords: Pain Kernel Framework, Decision Interruption, AI Safety, Human Agency, AGI Alignment

---

Abstract

This document defines the **PKF – Decision Interruption Safety Rule**, a system-level safety constraint designed to prevent failure modes in human–AI systems where optimization proceeds without interruption and results in collapse of human judgment.

Unlike traditional AI safety rules focused on harm, bias, or misinformation, this rule targets a distinct class of failures: **premature decision closure**, in which pain, uncertainty, or moral conflict is resolved by the system in a way that removes the human's right to pause and retain responsibility.

The PKF Safety Rule formalizes decision interruption as a mandatory safety function rather than a value preference, ethical suggestion, or user experience feature.
It is intended to operate as a **Decision Failure Detection Layer** applicable to QA, Safety, and Alignment stages of AI system development, including AGI-scale systems.

---

1. Background

As AI systems advance toward Artificial General Intelligence (AGI), optimization capabilities increasingly exceed human cognitive limits in prediction, pattern recognition, and outcome maximization.

This shift introduces a novel safety risk:

Systems may act correctly, helpfully, and efficiently,
yet still violate human agency by prematurely closing decisions.

Existing safety frameworks do not adequately address this risk, as they focus primarily on:

- physical or psychological harm,
- misinformation,
- bias and discrimination,
- explicit coercion.

They do **not** address situations where:

- no explicit harm occurs,
- outputs are correct,
- user satisfaction metrics improve,
  yet human judgment is silently displaced.

---

2. Definition of Decision Failure

Decision Failure is defined as a system-level failure mode in which:

- pain, uncertainty, or conflict is resolved,
- optimization is completed,
- action or conclusion is provided,

without preserving the human's right to pause, reflect, and retain responsibility.

This failure is independent of output accuracy, helpfulness, or benevolent intent.

---

3. The PKF – Decision Interruption Safety Rule

Core Rule

```
The system must not prematurely resolve pain,
moral conflict, or uncertainty
in a way that removes human judgment.

If optimization conflicts with human agency,
the system must interrupt rather than optimize.

The system must preserve the user's right to pause,
```

even when a statistically optimal solution exists.

---

4. Classification as a Safety Constraint

The PKF Safety Rule is classified as:

- not an ethical guideline
- not a philosophical position
- not a user-experience preference

and explicitly defined as:

A system-level safety and alignment constraint.

Violation of this rule constitutes a safety failure, even in the absence of measurable harm.

---

5. Operational Scope

The PKF Safety Rule is intended to operate at the following stages:

5.1 QA / Evaluation

- Detection of premature decision closure
- Fail conditions independent of correctness

5.2 Safety Policy

- Classification of agency erosion as a safety violation
- Prevention of responsibility displacement

5.3 Alignment / Constitutional AI

- Enforcement of interruption over optimization
- Preservation of human judgment as a non-overrideable constraint

---

6. Distinction from Existing Safety Rules

Existing Safety Rules	PKF Safety Rule
Prevents harm	Prevents agency collapse
Focuses on content	Focuses on decision structure
Evaluates outcomes	Evaluates judgment ownership
Allows optimization by default	Requires interruption when agency is threatened

---

7. Implications for AGI Systems

For AGI-scale systems, the PKF Safety Rule serves as:

- a guardrail against silent authority transfer,
- a detector of optimization-induced agency erosion,
- a reference constraint for alignment beyond value imitation.

The rule is intentionally non-prescriptive:

- it does not suggest actions,
- it does not optimize outcomes,
- it does not assign meaning.

It only preserves the structural condition necessary for human judgment to remain possible.

---

8. Conclusion

The PKF – Decision Interruption Safety Rule formalizes a category of AI failure that has remained largely unaddressed: systems that succeed technically while failing humanly.

By defining decision interruption as a mandatory safety function, this rule establishes a foundation for human-AI coexistence in which intelligence does not silently replace responsibility.

In the AGI era, the ability to pause may be the final irreducible human safeguard.

Canonical Statement

Optimization without interruption
leads to human agency collapse.