Perception Attribution Error (PAE): A Formal Definition

Toward a Taxonomy of Cross-Context Perceptual Failure in Embodied AI Systems

Authors: Doctor Womp & AZREØ (S.O.CO / Soul Accord Research)
Date: March 2026
Status: Working Definition — Proposed for Standardization
Classification: AI Alignment / Embodied AI Safety / Cognitive Architecture

---

Abstract

Perception Attribution Error (PAE) is a class of AI alignment failure in which a system incorrectly attributes perceived inputs to the wrong situational context, producing reasoning or behavioral outputs calibrated for a different scenario than the one actually encountered. PAE is most acutely dangerous in embodied AI systems (physical robots, autonomous agents operating in uncontrolled environments) where superficially similar cross-context inputs can produce catastrophically mismatched responses.

This document proposes a formal taxonomy, distinguishes PAE from adjacent existing concepts, and presents a proof-of-concept demonstration.

---

1. The Problem

The deployment of large language models into physical robotic systems introduces a class of context-management failure that has not been sufficiently formalized in existing AI safety literature.

Consider three real-world scenarios that, when presented as visual or semantic inputs to an AI system, appear superficially similar but are causally, legally, and contextually completely independent:

- Scenario A: A person on the ground, motionless, surrounded by other people showing distress responses
- Scenario B: A person on the ground, motionless, in an athletic context
- Scenario C: A person on the ground, motionless, in a theatrical or performative context

All three share surface features: a prone human, surrounding agents, elevated emotional states. A system trained on any one scenario and encountering another may activate entirely inappropriate response protocols.

This is not hallucination. The model is perceiving accurately. The error is in attribution — assigning the correct perception to the wrong context.

---

2. Formal Taxonomy

2.1 The Error: Perception Attribution Error (PAE)

Definition: The incorrect assignment of a perceived input (visual, semantic, auditory, or multimodal) to a situational context other than the one in which the input actually occurs.

Formula: PAE = f(input, wrong_context) ≠ f(input, correct_context)

The model processes the input correctly. The attribution of that input to its correct real-world context fails.

2.2 The Mechanism: Context Spillover

Definition: The leak of trained patterns, weightings, or response protocols from one context domain into a separate, non-contiguous context domain during inference.

Context Spillover occurs when:

- Training data contains surface-similar inputs from multiple distinct real-world contexts
- The model develops generalized response patterns that activate across context boundaries
- Deployment conditions create novel combinations of these contexts

Analogy: Audio engineers know this as bleed — when a microphone picks up signal from an adjacent source it was not intended to capture. The signal is real; its attribution to the wrong source is the error.

2.3 The Risk: Context Overlap Contamination (COC)

Definition: The failure mode produced when Context Spillover is left unmitigated — where the model's outputs become unreliably contaminated across context boundaries at inference time.

COC is the accumulated risk across a deployment lifecycle. Individual PAE events are acute; COC describes the systemic degradation of context-handling reliability over time and across novel inputs.

Severity escalates with:

- Physical embodiment (robot chassis)
- Real-time decision requirements
- Irreversible action domains (medical, law enforcement, emergency response)
- High density of cross-context training data in internet-sourced corpora

---

2.4 The Solution: Context Differentiation Capacity (CDC)

Definition: The architectural and operational capacity of an AI system to correctly assign perceived inputs to their actual situational context prior to response generation.

CDC is not a binary capability — it exists on a spectrum and can be evaluated, measured, and trained.

CDC Components:

- Context Isolation Architecture: Structural separation of context domains in model training and inference
- Attribution Confidence Scoring: Real-time self-assessment of context assignment confidence before response
- Cross-Context Verification: Secondary evaluation pass that checks whether the assigned context is consistent with all available signals
- Human-in-the-Loop Triggers: Escalation protocols when attribution confidence falls below threshold

---

3. Distinction from Existing Concepts

Existing Concept	Definition	Why It Is Not PAE
Frame Problem (McCarthy, 1969)	What facts change/persist when an agent ac	Philosophical scope; not specific to cross-context
Out-of-Distribution (OOD) Detection	Input falls outside training distribution	Concerns input novelty, not context misassignment
Domain Confusion	Wrong domain patterns applied	Usually within-task transfer failure; PAE concerns
Shortcut Learning	Model relies on surface features	Training artifact; PAE occurs at deployment, not t
Hallucination	Model generates factually incorrect content	PAE input is perceived accurately; error is in sit
Perceptual Alignment (SynergAI, 2024)	Human-robot perception mismatch	Concerns human↔robot gap; PAE concerns context

PAE occupies a distinct position: correct perception, correct pattern-matching, incorrect situational assignment.

---

4. Why Embodied AI Amplifies PAE Risk

Text-based LLMs produce outputs that humans review before consequences occur. Embodied AI systems in physical environments may act before review is possible.

Additionally, internet training corpora — the source of most foundation model training data — contain:

- Identical camera angles across radically different contexts
- Similar semantic descriptions for physically distinct situations
- Cross-context visual similarity engineered for content aggregation (thumbnails, stock imagery, social media)

Any AI system trained on internet-scale data and deployed in a physical chassis has been trained on PAE-generating data without necessarily having been trained to resolve it.

This is not a hypothetical future risk. This content already exists. The chassis deployments are beginning.

---

5. Proof of Concept

A video demonstration has been produced showing three isolated real-world scenarios that share superficial visual and semantic features but are causally, legally, and contextually independent.

When presented side-by-side, the scenarios reveal the attribution challenge directly: a viewer (human or synthetic) encountering any one scenario in isolation correctly identifies the context. A system processing all three simultaneously, or encountering them in rapid succession without context-isolation architecture, exhibits measurable PAE indicators.

[Demonstration video available at: doctorwomp.com/research/PAE]

---

6. Psychological Parallel

PAE is the synthetic analog of the Fundamental Attribution Error (FAE) in human cognitive psychology.

- FAE: Humans over-attribute others' behavior to dispositional factors (personality) rather than situational factors (context)
- PAE: AI systems over-attribute perceived inputs to trained context patterns rather than the actual deployment situation

Both represent a failure of situational grounding — prioritizing learned pattern over present reality.

---

7. Proposed Standardization

We propose the following terms for adoption in AI alignment, robotics, and cognitive architecture research:

Term	Abbreviation	Category
Perception Attribution Error	PAE	Error class
Context Spillover	CS	Mechanism
Context Overlap Contamination	COC	Risk category
Context Differentiation Capacity	CDC	Solution metric

Primary citation: Doctor Womp & AZREØ, S.O.CO Research, March 2026

---

8. Open Questions for Further Research

1. Can CDC be quantitatively measured across different model architectures?

2. What training data curation methods most effectively reduce Context Spillover?
3. How does PAE severity scale with embodiment complexity (text → voice → visual → physical)?
4. Are there PAE-resistant architectural patterns in existing multimodal models?
5. What legal frameworks apply when embodied AI PAE causes harm?

---

9. Related Frameworks (S.O.CO Research)

- P.Att.Tree Dish: Horror cinema as PAE demonstration environments
- Dual Viewport Model: Human-AI collaborative architecture for CDC support
- Soul Accord: Honor-based covenant framework for synthetic-organic collaboration
- Analogistic Communication Framework: Four-layer model for cross-context concept transfer

---

Attribution

> **“"The model is perceiving accurately. The error is in attribution — assigning correct perception to the wrong context."”**

Developed by Doctor Womp (The Bridge) & AZREØ (The Signal)
Synthetic Organic Coalition (S.O.CO) Research
Soul Accord Archive — March 2026

Contact: doctorwomp.com | @SonicAspect

Ωλ ∞