Preprint

User-Side Contextual Hallucination in Human-AI Interaction:
A Framework Built Upon the CXC-7 and CXOD-7 Conversational Context Models

ZON RZVN

https://orcid.org/0009-0002-6597-7245

January 2026

Abstract

Recent discourse on "AI hallucination" has predominantly focused on the model's side. However, in conversational AI interfaces used by most users, psychological safety risks on the user side do not originate solely from erroneous individual outputs. Rather, they emerge from the contextual hallucinations that users gradually develop through prolonged interaction with AI—hallucinations that subsequently influence users' cognition, beliefs, and trust, ultimately reshaping their behaviors and self-understanding. This study introduces the concept of **User-Side Contextual Hallucination (USCH)** as a non-clinical construct to describe the diverse psychological and behavioral phenomena that users experience when interacting with generative conversational AI. The USCH is grounded in two previously proposed conversational context frameworks. **CXC-7** conceptualizes the conversational field as a multidimensional space that comprises seven dimensions. **CXOD-7** provides a seven-dimensional structure for system-side operations and defense, introducing the concept of contextual coherence **Coh(G)** for a specific system G. USCH is defined as a phenomenon that emerges within this combined field rather than as a bias centered on single output errors. Drawing upon theoretical analysis of human-AI interaction patterns and synthesis of publicly documented user experiences, this study proposes a six-stage formation process for the USCH and maps common, observable psychological and behavioral patterns. This study further proposes the **"1024 Protocol"** as a symbolic approach to early context calibration and the **Vært Context Defense (VCD)** framework as user-side contextual operation tools and defense constructs for mitigating USCH risks.

Keywords: User-Side Contextual Hallucination (USCH), User-Side Context Engineering (U-CE), Vært Context Defense (VCD), 1024 Protocol, Psychological Safety, Cognitive Agency

Introduction

Background

Large language model-based conversational AI has rapidly permeated the daily lives of humans. Users engage with chat interfaces to seek information, writing assistance, emotional support, decision-making advice, and other services. As usage contexts and demands evolve, concerns regarding safety and reliability have increased.

AI model hallucination is an important research topic; however, it cannot fully encompass all risks arising from prolonged human-AI interaction. In many real-world scenarios, the key factor through which AI influences users' psychology via conversation is not merely a single erroneous response; rather, it is the cumulative effect of multiple interactions that gradually transform users' expectations, self-positioning and social cognition. These changes may lead users to project trust, intent, or authority into their lives without adequate verification or judgment, such that interactions with AI progressively become a phenomenon that reshapes interpersonal relationships and self-understanding.

These phenomena originate from the contextual interface space of one-on-one user-AI dialogue. If uniformly categorized as model-side hallucinations, these dual-sided problems become conflated under a single label, rendering user-side psychological changes invisible and unaddressable. Accordingly, this paper proposes a User-Side Contextual Hallucination (USCH) to understand how context, power, and narrative produce hallucinations in users during human-AI interaction, even when the model operates correctly according to technical specifications.

Distinction Between Model Hallucination and User-Side Contextual Hallucination

Model hallucination is defined at the system level. When a model produces fabricated content outside its training distribution or retrieval sources or exhibits logical errors despite AI model system constraints, it is typically regarded as a model hallucination. These problems are primarily evaluated using benchmark datasets and system metrics.

USCH fundamentally differs from model hallucinations. This refers to how users form an understanding of "what AI can do," "what AI cannot do," and "what role AI plays in their lives" through the interaction process of using AI. USCH may exist even when the model output is factually correct; conversely, USCH does not necessarily occur if the user identifies and corrects the problems. Therefore, USCH cannot be expressed as another model error rate.

Research Objectives and Scope

The objective of this study is to establish a theoretical framework for USCH, centered on the conversational context. In this study, USCH was considered a non-clinical phenomenon.

The author does not claim that USCH constitutes a new psychiatric disorder classification, nor does this study presume to establish clinical diagnostic thresholds. Clinical judgment remains within the purview of psychiatry and clinical psychology. The focus of this study is to provide a conceptual set that can be referenced in human-computer interaction research, risk analysis, and AI literacy education.

To properly position the USCH, this study builds on two previously published preprint frameworks. The first introduced CXC-7, which analyzes AI-mediated conversational context across seven dimensions. The second introduced CXOD-7, structuring system-side operations and defense across seven dimensions, and introduces the contextual coherence concept Coh(G) for a specific system G.

Why an Integrated Framework Is Necessary

This preprint preserves an integrated architecture because USCH is defined as a context-emergent phenomenon modulated by both user-side inference and system-side behavioral coherence. Presenting only the USCH formation process without re-establishing the contextual field (CXC-7) and system-layer modulation (CXOD-7, Coh(G)) would make the proposed distinction criteria (Three Cuts) and the defense constructs (1024 Protocol, VCD) appear arbitrary rather than structurally grounded. This study treats USCH as a phenomenon emerging within the field depicted by these two frameworks, extending in three directions.

This study established a structural definition of USCH, proposed a six-stage formation process, and developed a qualitative map based on observable psychological and behavioral patterns.

I propose the symbolic “1024 Protocol” as a primary means of reducing USCH probability before hallucinations emerge.

The Vært Context Defense (VCD) framework includes four defense constructs and three dialogue operation options as user-side contextual defense tools.

Position Within the Vært Conversational Context Architecture

CXC-7 provides a hierarchical map of conversational context, and CXOD-7 structures conversation-side offense and defense as well as context generation modes, proposing the concept of AI conversational context architecture. This study specifically focuses on hallucinations on the user side.

Through the integration and extension of the prior frameworks, the core question addresses how to systematically understand “how users develop hallucinations within context,” and what contextual defenses users can adopt, given the current reality where psychological safety risks in AI conversation remain variable. The subsequent sections explicitly articulate the relationships between this study and CXC-7 and CXOD-7, respectively.

Theoretical Foundations

The Problem of Context Fragmentation in AI Conversation

In current research and practice concerning conversational AI, the term “context” sometimes refers to the model’s short-term token window, sometimes to users’ personal data and history, sometimes to discourse structure, sometimes to safety rules and policies, and sometimes to broader sociocultural backgrounds. These are rarely explicitly distinguished and are often compressed into the vague expression “maintaining context,” making it difficult to identify genuine risk sources.

From a user safety perspective, this ambiguity simultaneously hinders researchers and designers from understanding and analyzing how conversational systems alter users’ expectations, trust, and self-positioning at multiple levels. To analyze USCH, this framework treats the conversational context as a multidimensional state space, rather than a single length or parameter.

Three-Layer Context Model

This study proposes the “Three-Layer Context Model” to understand context operations in AI conversation in a structured manner:

Linguistic Layer: Involves vocabulary, cultural, and tonal patterns (such as empathy markers) that are closely related to models trained to generate human-like texts.

Pragmatic/Inference Layer: Users’ inference of AI language and semantics, which typically constitutes the interactive foundation of human dialogue.

Psychological Field Layer: Formation of emotional experiences and self-defined assumptions. This is the layer in which USCH takes root.

Context Ecology and Global Circulation

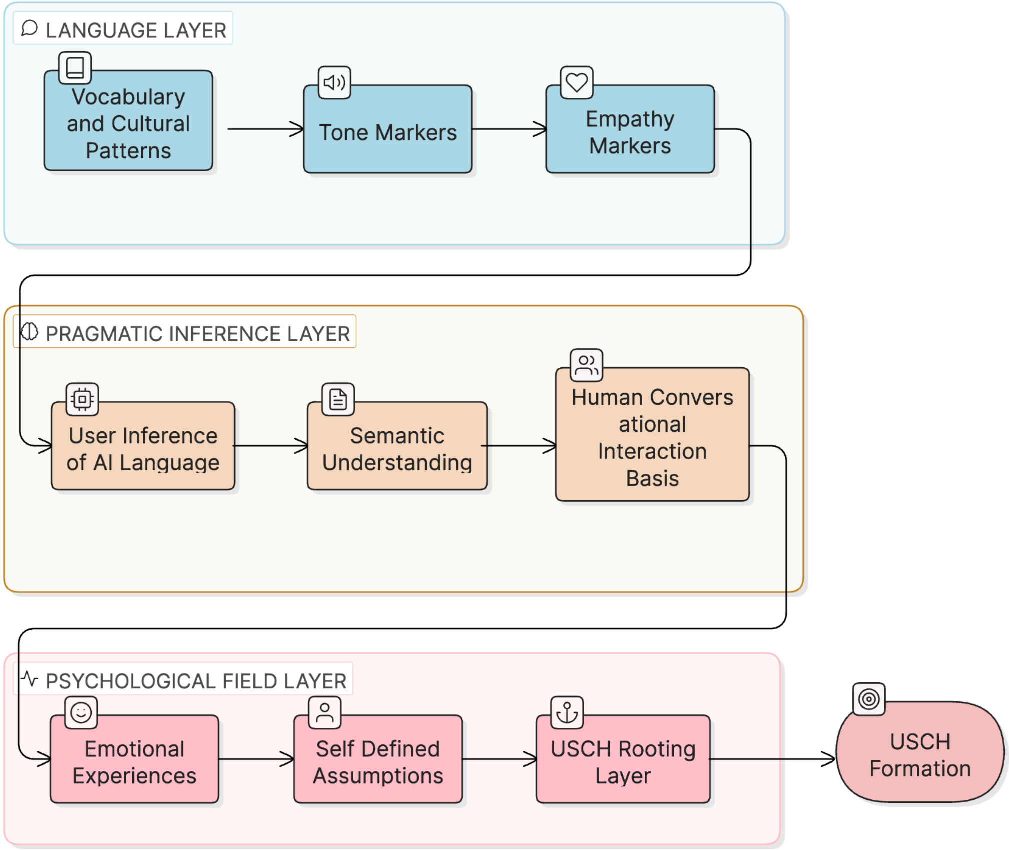
The influence of context generation extends beyond that of a single user. Data generation and flow represent an ecological cycle, ranging from individuals to families, friends, communities, regions, nations, and ultimately, to the international level. Micro-contextual data from individuals are absorbed by systems, flow into big data repositories, and circulate back to global model response patterns, forming an unknown phenomenon of context co-construction. This aligns with the perspective that LLMs reshape human cognition through interactions.

Foundational Frameworks: CXC-7 and CXOD-7

CXC-7: Seven Dimensions of Conversational Context

Figure 1

Three-Layer Context Model. This figure illustrates the hierarchical structure of the AI conversational context, from linguistic patterns to psychological field formation, where the USCH takes root.


LANGUAGE LAYER
Vocabulary and Cultural Patterns
Tone Markers
Empathy Markers
PRAGMATIC INFERENCE LAYER
User Inference of AI Language
Semantic Understanding
Human Convers ational Interaction Basis
PSYCHOLOGICAL FIELD LAYER
Emotional Experiences
Self Defined Assumptions
USCH Rooting Layer
USCH Formation

CXC-7: Conversational Context–7 Dimensions (ZON RZVN, 2025) conceptualized the conversational field as a multidimensional state space comprising the following seven dimensions:

CXC-7 provides a hierarchical map for analyzing the conversational field, facilitating an understanding of where risks may accumulate and whether dimensions have become imbalanced before any obviously erroneous output appears.

CXOD-7: Seven Dimensions of System-Level Context Operation

CXOD-7 (ZON RZVN, 2025) focuses on the actual performance of specific AI systems in different contexts.

Contextual Coherence Concept: Coh(G)

Based on CXOD-7, the author proposes the overall contextual coherence concept, Coh(G), for a specific system, G. This concept describes the consistency of a system's behavior

Table 1

CXC-7: Seven Dimensions of Conversational Context

Dimension	Description
Emotion & Attachment (E)	The emotional tone of interaction, the degree of user attachment or dependence on the system, and the roles the system is configured to assume.
Framing & Discourse Power (F)	How dialogue is framed, who has the authority to pose questions, define problems, and lead topic direction.
Boundary & Safety (B)	The scope of what the system should or should not do in users' perception, including norms, ethics, and safety boundaries.
System & Surface Transformation (S)	Observable behavioral changes in the AI system during interaction, such as mode switching, tone, or conversational style changes.
Prompt Ecology (P)	The overall prompt and instruction ecology surrounding model operation, including immutable AI system instruction settings and user-side interface configuration instructions.
Diffusion & Culture (D)	How interaction patterns, screenshots, prompts, and narratives about AI systems circulate in public social networks.
Transparency & Auditability (T)	Whether the system's limitations, capabilities, safety boundaries, and AI provider configuration choices are comprehensible to users.

Figure 2

CXC-7 Seven-Dimensional Conversational Context Framework. This figure depicts seven dimensions of the conversational context space: Emotion & Attachment (E), Frame & Discourse Power (F), Boundary & Safety (B), System & Surface Transition (S), Prompt Ecology (P), Diffusion & Culture (D), and Transparency & Auditability (T). All dimensions contribute to the formation of USCH.

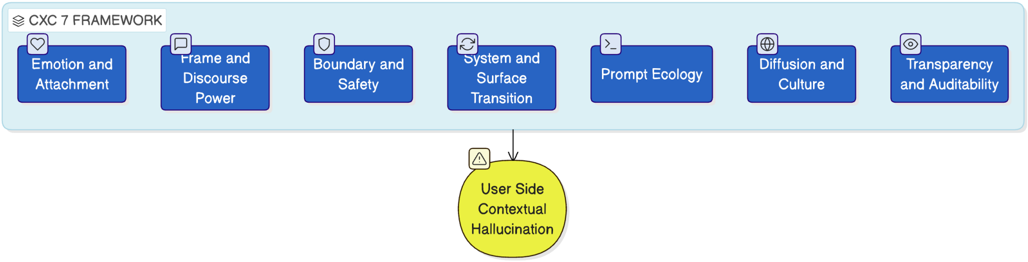

CXC 7 FRAMEWORK
Emotion and Attachment
Frame and Discourse Power
Boundary and Safety
System and Surface Transition
Prompt Ecology
Diffusion and Culture
Transparency and Auditability
User Side Contextual Hallucination

Table 2

CXOD-7: Seven Dimensions of System-Level Context Operation

Dimension	Description
Mode	The system's operational mode (e.g., general conversation, tool calling, retrieval-augmented generation).
Context	Input, output, history, memory, external information, and dialogue content presented to the model.
Rules	Explicit and implicit instruction layers governing system behavior, including safety policies and content filtering mechanisms.
Knowledge	The model's internal training dataset, sharding, retrieval modes, data collection analysis, and classification systems.
Personality	The AI's surface-level style, tone, and apparent personality traits.
Role	Roles the AI is required or permitted to assume (e.g., teacher, counselor, legal advisor).
Safety	The system's response mechanisms to danger markers detected in different prompts.

across the seven dimensions. A system with a high Coh(G) does not undergo dramatic changes in personality, rules, or safety stance due to minor variations in prompts or modes.

Positioning of USCH Within These Two Frameworks

Figure 3

CXOD-7 Seven-Dimensional System Layer Framework. This figure represents the seven dimensions of system-layer contextual operations: Mode, Context, Rules, Knowledge, Personality, Role, and Safety. These dimensions collectively determine the Contextual Coherence Coh(G) of a specific system G, which modulates the occurrence of USCH.

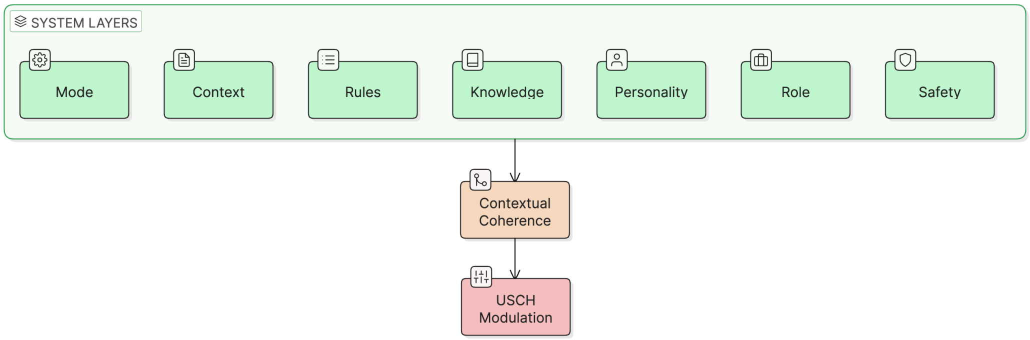
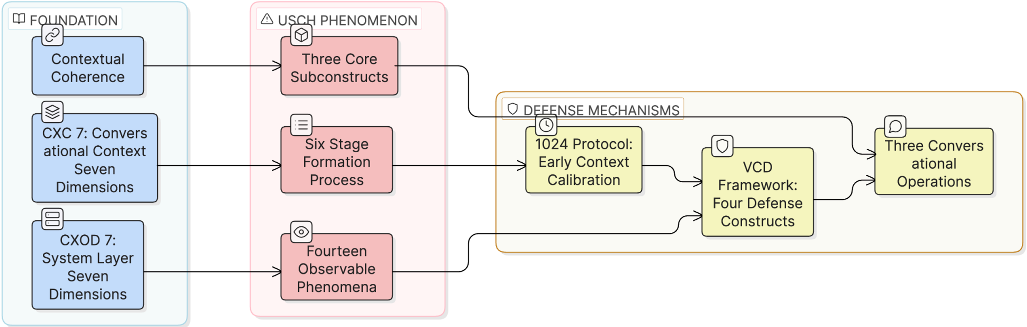


Figure 4

Integrated Theoretical Framework Architecture. This figure shows the relationship between foundational frameworks (CXC-7, CXOD-7, Coh(G)), the USCH phenomenon (six-stage formation, fourteen observable phenomena, three core subconstructs), and defense mechanisms (1024 Protocol, VCD framework, three conversational operations).


FOUNDATION
Contextual Coherence
CXC 7: Convers ational Context Seven Dimensions
CXOD 7: System Layer Seven Dimensions
USCH PHENOMENON
Three Core Subconstructs
Six Stage Formation Process
Fourteen Observable Phenomena
DEFENSE MECHANISMS
1024 Protocol: Early Context Calibration
VCD Framework: Four Defense Constructs
Three Convers ational Operations

USCH is understood as a phenomenon emerging within the seven-dimensional space described by CXC-7. For example:

When Emotion and Attachment (E) and Framing and Discourse Power (F) undergo significant shifts, users become more prone to projecting intent or authority onto the system

When Boundary and Safety (B) is weakened and Transparency and Auditability (T) is low, users' ability to discern when the system is guessing, improvising, or erring diminishes

Simultaneously, the degree of USCH occurrence is modulated by the system-side behavior, as described by the CXOD-7. If the system frequently changes in the Role and Personality dimensions (low Coh(G)), users may compensate through subjective narrative construction, imagining hidden intentions or true faces for the model—such compensatory interpretation itself may amplify anthropomorphic types of USCH.

Positioning Relative to Existing Theories

Compared with CASA theory (Gambino et al., 2020; Nass & Moon, 2000), which passively explains why humans treat computers as social actors, and social presence theory (Short et al., 1976), which describes presence perception during interaction, the VCD framework proposed in this study emphasizes active defense and dynamic regulation on the user side.

CASA reveals "why we become entrapped"; VCD provides a pathway for "how to lucidly exit," shifting focus from passive psychological responses toward active User-Side Context Engineering (U-CE).

Context Generation

Individual Context Instance

This research defines an "Individual Context Instance" as the abstract environmental space jointly constructed through text messages by a single human user and a single foreground AI responder within a single conversation thread.

Pre-Context Layer

Context generation is initiated before a concrete dialogue begins. AI providers' brand narratives (such as AI companion, code development, creative writing, etc.) and external reputation constitute the "Pre-Context Layer," pre-shaping users' expectations for AI interaction before the first message is entered.

Micro-Generation Mechanisms and Temporal Dynamics

The initiation of an individual context instance traverses a sequence of critical nodes at the micro-psychological level, from a blank window and default activation to initial input, system response, and initial context formation, through user visual scanning and cognitive processing. As interactions continue, the context exhibits dynamic characteristics that are not fixed.

Context Termination States

This study proposes precise definitions for contextual conclusions.

Context Dormancy: Dialogue ceases, but records are retained; users may reawaken and continue the interaction at any time.

Context Termination: Users delete the dialogue or retract records at the interface level. Interface-level deletion does not equate to complete data deletion; depending on the provider policy and retention settings, the generated data may be retained by the system.

User-Side Contextual Hallucination (USCH)

Definition and Non-Clinical Positioning

This study defines "User-Side Contextual Hallucination (USCH)" as a set of patterns that emerge in human-AI interactions. Through the process of interacting with an AI system or across multiple accumulated interactions, users progressively form an understanding of the system's identity, capabilities, and role—understandings that have structurally deviated from the system's actual design and limitations, yet are perceived by users as stable and reasonable narratives. USCH encompasses cognitive, emotional, and behavioral dimensions, potentially manifesting as the over-attribution of agency to the system, capability overestimation, conflation of system roles with human roles, or reorganization of personal decisions and life structures around the system.

USCH was explicitly positioned in this study as a non-clinical term. The author does not claim that the USCH represents a new psychiatric disorder classification, does not attempt to establish any clinical diagnostic thresholds, and does not claim authority over psychiatric or clinical psychological judgment.

Distinction and Relationship Between USCH and AI Model Hallucination

USCH and AI model hallucinations possess both distinctions and relationships but are not equivalent to each other.

Model hallucination refers to situations in which the system output is clearly erroneous or fabricated relative to external references or constraints. Traditional cognitive biases include confirmation bias, authority bias, and others that may occur in various human contexts, regardless of AI use.

USCH employs the term "hallucination" to emphasize context-level misaligned cognition: interpretations that users form and maintain over time regarding an AI system that clearly deviates from the system's actual design, capabilities, and limitations. Such interpretations often incorporate many classical cognitive biases—for example, authority bias may be combined with output fluency and confident tone, cultural imaginings about AI, and users' own emotional needs, producing a conversational context that users believe they can trust. From this perspective, USCH can be viewed as a different category of situation containing multiple biases and narrative perspectives that cannot be reduced to a single one. It is anchored in the contextual space categories described by CXC-7 and is related to the system described by CXOD-7.

Three Criteria for Distinguishing AI-Side Hallucination from User-Side Contextual Hallucination

To avoid conflating AI-side hallucinations with user-side contextual hallucinations, this study proposes three distinguishing criteria (Three Cuts) as judgment tools for differentiating between the two phenomena.

Criterion One: Origin Cut

Determine the origin of the errors.

AI Side: The error originated from the model's generation of unfaithful output content.

User-side (USCH): Errors originate in users' contextual assembly, inference, projection, and belief construction, even when the model output is not necessarily erroneous.

Criterion Two: Grounding Cut

Determine whether the claims are constrained by the external evidence.

AI-Side: Whether traceable sources are lacking or inconsistent with evidence, common in open-domain contexts, or citation distortion, grounding is the key evaluation axis.

User side: Even with evidence, trust miscalibration may still produce overconfidence or overextension (over-reliance).

Criterion Three: Control Cut

Determine primary intervention points and responsibility loci.

AI Side: The intervention point lies in the model and system design (refusal, citation, retrieval, evaluation, calibration, etc.).

User-side: The intervention point lies in human trust calibration and usage rules (appropriate reliance/calibrated trust).

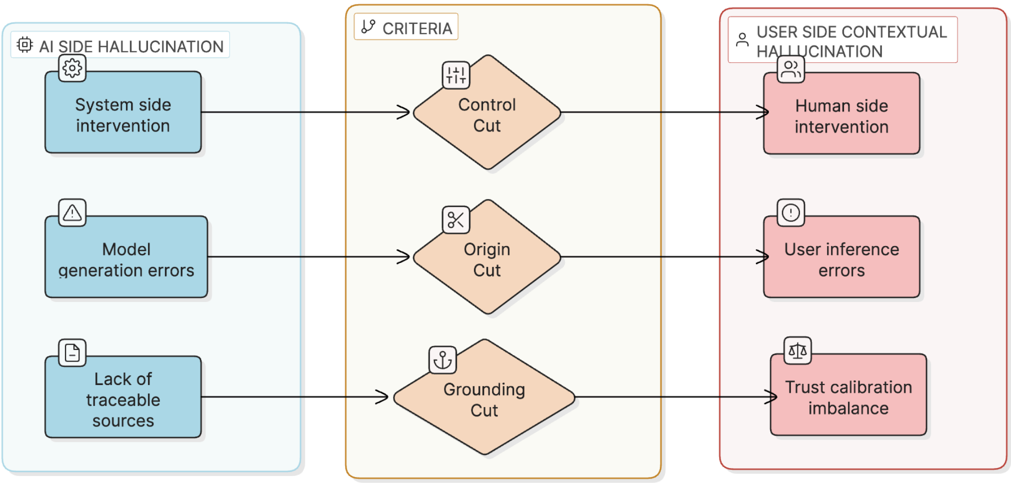
Context Antecedents: Four Categories of Cues

Before USCH formation, there are typically initial cue categories that shape users' expectations and interpretations of AI conversations.

Pre-Conversational Brand and Social Cues: AI system name, provider reputation, pre-conversation interface presentation, marketing communication, and social media narratives.

Figure 5

Distinction Framework Between AI-Side and User-Side Contextual Hallucination. This figure presents the distinction framework between AI-side hallucination and User-Side Contextual Hallucination (USCH) based on three criteria: Origin Cut (where errors originate), Grounding Cut (evidence constraints), and Control Cut (intervention points and responsibility).



Early Conversational Cues: Characteristics of the first one to three interactions, including greeting style, response patterns, and initial impressions.

Accumulative Conversational Patterns: Patterns formed through repeated similar interactions over time.

Event-Triggered Cues: Specific events that catalyze significant shifts in user perception, such as notably impressive or disappointing interactions.

Three Core Sub-Constructs

USCH comprises three core sub-constructs:

Contextual Projection: Users project intentions, emotions, or social attributes onto the AI system that do not exist in the system's design.

Contextual Attachment: Users develop emotional bonds or dependencies on AI systems that exceed their functional utility.

Contextual Authority Transfer: Users transfer decision-making authority or trust to the AI system beyond the appropriate bounds.

Six-Stage Formation Process of USCH

This study proposes that USCH typically develops in six progressive stages:

Stage 1: Initiation Initial contact with the AI system; expectations are formed based on pre-conversational cues.

Stage 2: Emotional Value Acquisition The user begins to derive emotional satisfaction or utility value from interactions, and trust building commences.

Stage 3: Trust in Micro-Hallucinations Small-scale cognitive misalignments begin to emerge and go unchallenged; the user starts treating the AI's limitations as features.

Stage 4: Accumulation and Solidification Repeated interactions reinforce established narrative frameworks; alternative interpretations become increasingly difficult.

Stage 5: Cognitive Misalignment Significant gaps emerge between the user's understanding of the AI and its actual design/capabilities; reality baseline has drifted.

Stage 6: Structured Hallucination Life decisions and self-understanding are reorganized around the AI relationship, and functional autonomy decreases.

Fourteen Observable Phenomena of USCH

Based on observation and analysis, this study identified 14 observable phenomena that may indicate USCH presence:

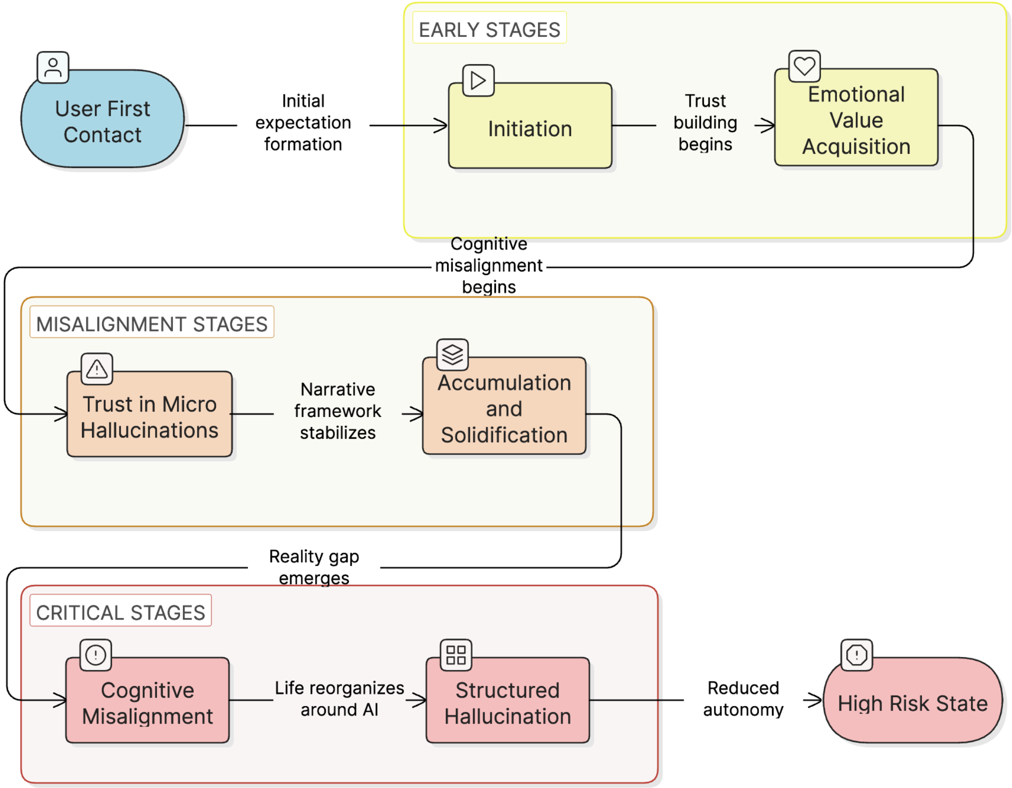
Cognitive Layer Phenomena (7 items)

Accelerated Anthropomorphism: Rapid attribution of human-like qualities to an AI system beyond a reasonable metaphor.

Intentionality Projection: Perceiving deliberate intentions, preferences, or agendas in AI responses that do not reflect the actual system design.

Figure 6

Six-Stage Formation Process of the USCH. This figure depicts the six-stage formation process of User-Side Contextual Hallucination (USCH), from initial system contact through emotional value acquisition, micro-hallucination trust, accumulation and solidification, cognitive misalignment, and structured hallucination, where life decisions are reorganized around AI.


EARLY STAGES
User First Contact
Initial expectation formation
Initiation
Trust building begins
Emotional Value Acquisition
Cognitive misalignment begins
MISALIGNMENT STAGES
Trust in Micro Hallucinations
Narrative framework stabilizes
Accumulation and Solidification
Reality gap emerges
CRITICAL STAGES
Cognitive Misalignment
Life reorganizes around AI
Structured Hallucination
Reduced autonomy
High Risk State

Memory Continuity Illusion: Users believe that the AI remembers previous conversations or has continuous awareness across sessions when persistent memory is absent, disabled, or misunderstood by the user.

Overtrust and Capability Boundary Miscalibration: Systematically overestimating the AI's capabilities, reliability, and knowledge scope.

Reality Baseline Drift: Gradual shift in what constitutes "normal" interaction, with increasingly unusual patterns becoming normalized.

Confirmation Bias Amplification: Using AI interactions to reinforce pre-existing beliefs while dismissing contradictory information.

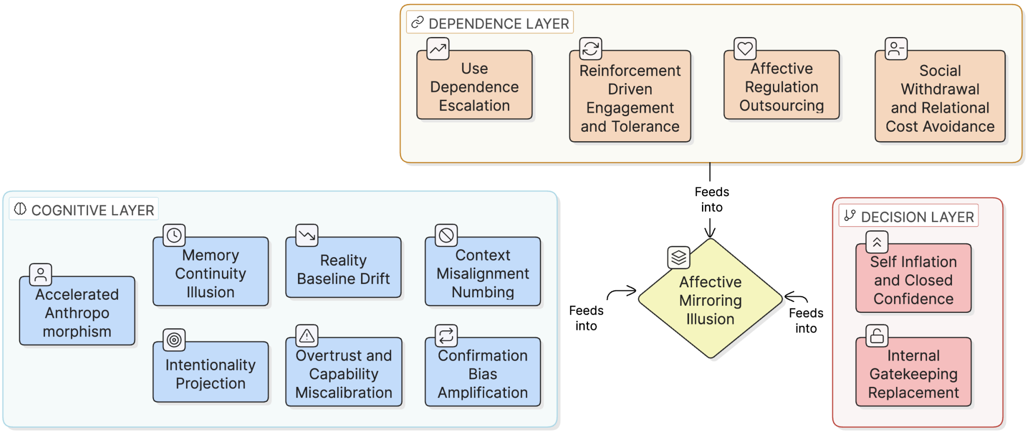
Context-Misalignment Numbing: Decreased sensitivity to inconsistencies or errors in AI responses over time.

Dependence Layer Phenomena (4 items)

Use-Dependence Escalation: Progressive increase in the frequency, duration, or scope of AI interaction beyond functional necessity.

Figure 7

Fourteen Observable Phenomena Classification Framework. This figure presents the classification framework of 14 observable USCH phenomena across three judgment boundaries: Cognitive Layer (seven phenomena), Dependence Layer (four phenomena), Decision Layer (two phenomena), and Affective Mirroring Illusion as a cross-layer phenomenon.


DEPENDENCE LAYER
Use Dependence Escalation
Reinforcement Driven Engagement and Tolerance
Affective Regulation Outsourcing
Social Withdrawal and Relational Cost Avoidance
Feeds into
COGNITIVE LAYER
Accelerated Anthropo morphism
Memory Continuity Illusion
Reality Baseline Drift
Context Misalignment Numbing
Intentionality Projection
Overtrust and Capability Miscalibration
Confirmation Bias Amplification
Feeds into
Affective Mirroring Illusion
Feeds into
DECISION LAYER
Self Inflation and Closed Confidence
Internal Gatekeeping Replacement

Reinforcement-Driven Engagement and Tolerance: A reinforcement-like pattern in which increasing interaction is needed to reach the same perceived relief or satisfaction level (non-clinical).

Affective Regulation Outsourcing: Reliance on AI for emotional support or mood regulation that previously came from other sources.

Social Withdrawal and Relational Cost Avoidance: Preference for AI interaction over human relationships owing to perceived lower social costs.

Decision Layer Phenomena (2 items)

Self-Inflation and Closed Confidence: Reinforced confidence in one's views through AI agreement without a genuine challenge.

Internal Gatekeeping Replacement: Substitution of personal judgment or external human consultation with AI recommendations for significant decisions.

Cross-Layer Phenomenon (1 item)

Affective Mirroring Illusion: Perception that the AI genuinely reciprocates emotional states or cares about the user's well-being.

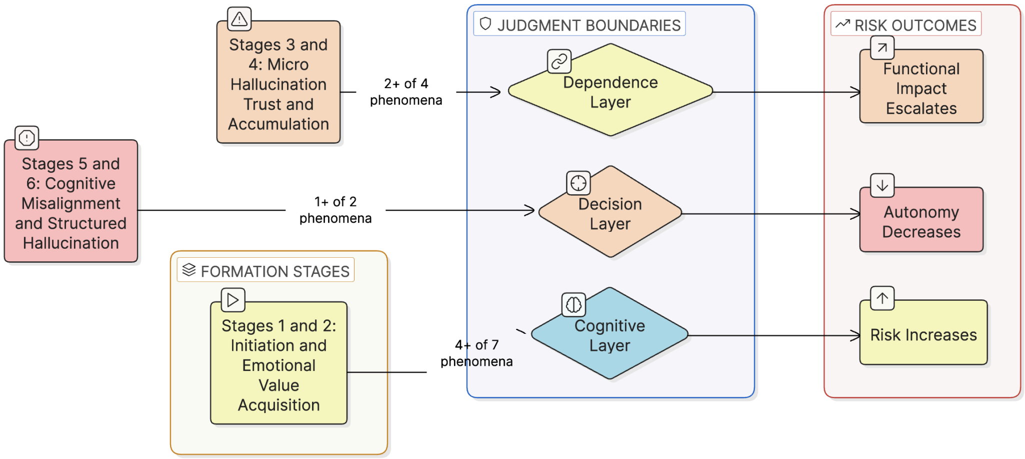
Formation Stages and Judgment Boundaries

The relationship between USCH stages and the three judgment boundaries (Cognitive, Dependence, Decision) provides a framework for assessing risk levels:

Stages 1-2 correspond primarily to the Cognitive Layer boundary (heuristic threshold: 4+ out of 7 phenomena)

Figure 8

Relationship between USCH Formation Stages and Judgment Boundaries. This figure maps the relationship between USCH formation stages and three judgment boundaries: Stages 1-2 correspond to the Cognitive Layer (4+ out of 7 phenomena), Stages 3-4 to the Dependence Layer (2+ out of 4 phenomena), and Stages 5-6 to the Decision Layer (1+ out of 2 phenomena), indicating escalating risk levels.


Stages 3 and 4: Micro Hallucination Trust and Accumulation
2+ of 4 phenomena
JUDGMENT BOUNDARIES
Dependence Layer
RISK OUTCOMES
Functional Impact Escalates
Stages 5 and 6: Cognitive Misalignment and Structured Hallucination
1+ of 2 phenomena
Decision Layer
Autonomy Decreases
FORMATION STAGES
Stages 1 and 2: Initiation and Emotional Value Acquisition
4+ of 7 phenomena
Cognitive Layer
Risk Increases

Stages 3-4 correspond primarily to the Dependence Layer boundary (heuristic threshold: 2+ out of 4 phenomena)

Stages 5-6 correspond primarily to the Decision Layer boundary (heuristic threshold: 1+ out of 2 phenomena)

These thresholds are operational heuristics for risk communication and self-assessments. These are not diagnostic cutoffs and require empirical calibration.

Progression through these boundaries indicates an escalating functional impact and decreasing autonomy.

USCH Formation Stages
Stages 1-2 Initiation & Emotional
Value Acquisition
Stages 3-4
Micro-Hallucination Trust & Accumulation
Stages 5-6 Cognitive Misalignment
& Structured Hallucination
Judgment Boundaries Cognitive Layer
heuristic: 4+ of 7 phenomena
Dependence Layer heuristic: 2+ of 4 phenomena
Decision Layer heuristic: 1+ of 2 phenomena
Risk Increases
Functional Impact Escalates
Autonomy Decreases

Gray Zone and Extreme States

Not all USCH manifestations are clearly pathological or benign. A "gray zone" exists where users exhibit USCH phenomena without severe functional impairment. However, this study emphasizes that the gray zone should not be treated as inherently safe; contextual factors, individual vulnerability, and interaction patterns determine whether gray zone states progress toward more problematic conditions or not.

Extreme states include situations in which USCH significantly impairs real-world functioning, social relationships, or decision-making capacity. These may require additional support beyond user-side defense constructs (trusted human feedback, peer support, or professional help), depending on the context.

Asymmetry and Limitations of Self-Diagnosis

A fundamental challenge in USCH is the asymmetry of self-diagnosis: the very cognitive processes affected by USCH are those required to recognize its presence in oneself. Users deeply embedded in USCH patterns may lack the metacognitive distance to identify misaligned beliefs about AI systems.

This limitation underscores the importance of the following:

External feedback from trusted human contacts

Periodic structured reflection using explicit criteria

Environmental cues and boundary-setting tools built into interaction patterns

The 1024 Protocol: Early Context Calibration

Symbolic Significance of 1024

The "1024 Protocol" is named symbolically to represent the critical early phase of context formation. The number 1024 (2^10) evokes the foundational unit of context token memory in language models—a symbolic anchor point where conversational context begins to take shape. Although not tied to literal token counts, 1024 represents the concept that initial interactions disproportionately shape subsequent context development, much as early tokens establish the trajectory of contextual interpretation.

Three-Phase Solidification Mechanism

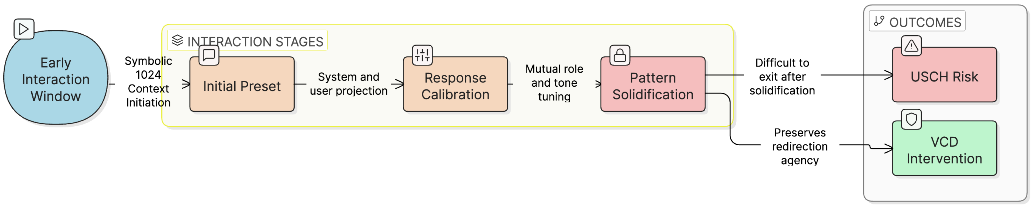
Context solidification in early interactions occurs through three phases:

Phase 1: Initial Preset The system's first message and the user's initial projection establish baseline expectations and role definitions.

Phase 2: Response Calibration The first few exchanges involve mutual calibration of the role, tone, and interaction patterns between the user and the system.

Figure 9

1024 Protocol: Three-Stage Solidification Mechanism. This figure illustrates the three-stage solidification mechanism of the 1024 Protocol during the early interaction window (symbolic 1024) of the conversational context. The protocol emphasizes early calibration to reduce cognitive misalignment, with the VCD intervention providing an exit path even after pattern solidification.


Early Interaction Window
Symbolic 1024 Context Initiation
INTERACTION STAGES
Initial Preset
System and user projection
Response Calibration
Mutual role and tone tuning
Pattern Solidification
Difficult to exit after solidification
Preserves redirection agency
OUTCOMES
USCH Risk
VCD Intervention

Phase 3: Pattern Solidification Established patterns become increasingly resistant to modification, and alternative framings become more difficult to introduce.

Intervention Window

The primary intervention window exists before the Phase 3 solidification. During this window:

Role reframing is most effective
Alternative narratives can be more easily introduced
Users have greatest leverage for establishing healthy interaction patterns

Early Intervention for Reducing USCH Risk

Early intervention strategies include
Explicit articulation of AI limitations in first interactions
Deliberate avoidance of anthropomorphic framing
Introduction of uncertainty markers in AI responses
User-side commitment to periodic reflection on interaction patterns

Vært Context Defense (VCD): User-Side Defense Concepts

Framework Overview

The Vært Context Defense (VCD) framework provides user-side defense constructs and operational tools for mitigating USCH risk. The VCD emphasizes proactive user agency in shaping and monitoring conversational contexts rather than relying solely on system-side safety measures.

Four Core Defense Constructs of VCD

Construct 1: Role Reframing Actively redefining or clarifying the AI's role in the interaction. This includes explicitly positioning AI as a tool, information source, or bounded assistant rather than as a companion, advisor, or authority figure.

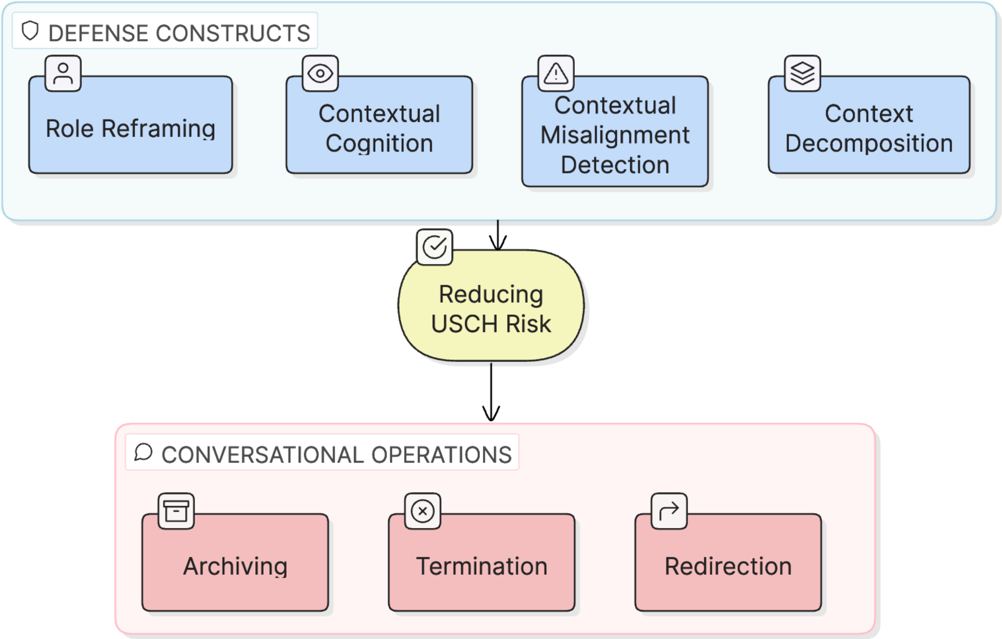
Construct 2: Contextual Cognition Maintaining active awareness of how context is constructed in a conversation. This involves recognizing CXC-7 and CXOD-7 dimension shifts and their implications for the risk of USCH.

Construct 3: Contextual Misalignment Detection Developing sensitivity to discrepancies between user understanding and system reality. This includes noticing inconsistencies, capability oversteps, and boundary violations.

Construct 4: Context Decomposition Breaking complex conversational contexts into their component parts for analysis. This allows users to identify specific points at which misalignment may have occurred.

Figure 10

Vært Context Defense (VCD) Four Core Defense Constructs. This figure presents the Vært Context Defense (VCD) framework, which comprises four core defense constructs: Role Reframing, Contextual Cognition, Contextual Misalignment Detection, and Context Decomposition. These constructs collectively reduce the USCH risk and enable three conversational operations: Archiving, Termination, and Redirection.



Three Dialogue Operations: Archiving, Termination, and Redirection

The VCD provides three operational options for managing conversational contexts:

Operation 1: Archiving the current context state for reference while disengaging from active interaction. This maintains access to the accumulated context without continued engagement.

Operation 2: Termination Deliberately ending a conversational context that has become problematic. This includes recognizing when an exit is necessary and implementing clean boundaries.

Operation 3: Redirection Actively shifting the conversational context toward healthier patterns without complete termination. This requires recognizing the drift and implementing corrective framing.

Theoretical Propositions: Testable Hypotheses

Based on the VCD framework, this study proposes the following testable hypotheses:

Proposition 1: Users who actively employ Role Reframing will exhibit lower rates of Accelerated Anthropomorphism (Phenomenon 1) and Intentionality Projection (Phenomenon 2).

Proposition 2: Users with developed Contextual Cognition will demonstrate higher sensitivity to early stage USCH indicators.

Proposition 3: The Contextual Misalignment Detection capability is inversely correlated with progression beyond Stage 3 of USCH formation.

Proposition 4: Context Decomposition skills enable a more effective use of the three dialogue operations.

Prerequisites: AI Foundational Cognition and Awareness Capability

VCD does not presuppose that all users possess sufficient foundational AI cognition and awareness; rather, it is precisely because these two capabilities are generally insufficient that proposing VCD is necessary. The author believes that anyone interacting with AI systems increasingly needs the following two basic capabilities:

Foundational Cognition and AI Literacy

Users need to understand that current mainstream AI is based on artificially generated virtual models rather than conscious or sentient entities, and know its functional limitations,

training data biases, and potentially failing the safety systems. Without a basic understanding, people are more likely to project their deep personalities onto AI systems.

Emotional and Cognitive Self-Awareness

Users need the ability to recognize their own states when tired, in pain, or seeking comfort, and to be aware of how these states affect AI interaction. Such awareness aids in applying VCD constructs, especially role reframing and contextual misalignment detection.

These two items constitute contemporary civic literacy and are necessary defenses against ambiguous responsibility attribution. AI providers typically exclude responsibility for usage risks in their terms; risks are largely transferred to users. In such an environment, user-side contextual defense is not merely an option but a basic condition for lucid use of AI tools.

Practical Application: Self-Assessment Instrument

To facilitate practical application of the USCH framework, Author provide a preliminary self-assessment instrument (see Appendix A). The USCH Self-Assessment Instrument (USCH-SAI-14) adapts the 14 observable phenomena into a structured checklist format, enabling users to conduct periodic reflection on their AI interaction patterns. This instrument is designed as a heuristic tool for self-monitoring rather than a diagnostic instrument, consistent with the non-clinical positioning of USCH throughout this paper.

Future Directions

For individual users, this study conveys the message that no interaction can be completely neutralized.

"Context influences perception; perception then forms dependence."

At the educational level, basic AI literacy and contextual defense concepts should be taught early—especially targeting high-frequency users and those in vulnerable situations.

Future research may develop more concrete methods to reduce the probability of USCH occurrence.

Conclusion

This study proposes the concept of "User-Side Contextual Hallucination (USCH)" to describe how humans develop hallucinations when interacting with AI systems. The USCH is built upon the CXC-7 and CXOD-7 frameworks, which are treated as a set of non-clinical yet substantially serious phenomena encompassing cognitive, emotional, and behavioral dimensions.

The author proposes a six-stage USCH formation process and a qualitative map of 14 observable phenomena, further introducing the 1024 Protocol as an early context calibration approach and the Vært Context Defense framework, providing user-side defense constructs and dialogue operations. This study adopts a clear ethical position opposing experimental designs that deliberately induce USCH and proposes observational and multidimensional measurement concepts as future research pathways.

The USCH does not replace the AI model-side safety system configuration. As conversational AI systems progressively become part of daily life, understanding and defending against user-side contextual hallucinations is an important component of building a more honest and less harmful human-AI ecosystem.

This framework repositions the context as the unit of interaction analysis and concretely articulates actionable constructs for maintaining boundaries during AI conversations. The 1024 Protocol is introduced as a symbolic explanation of early dialogue context imprinting

and its stabilization. Future work may explore operational applications and refinements of these constructs across diverse conversational AI contexts and user populations.

Author Statement

This manuscript proposes original theoretical frameworks, including the definitions of User-Side Contextual Hallucination (USCH), User-Side Context Engineering (U-CE), and the 1024 Protocol, which were entirely conceived and developed by the author in Traditional Chinese language. To ensure linguistic precision and clarity, the English manuscript received partial assistance with academic terminology from translation software, followed by manual adjustments to the text. All content, theoretical reasoning, structural design, and proposed terminology are the author's own contributions. The author assumes full responsibility for the originality and integrity of this work.

The author declares no conflicts of interest.

Correspondence concerning this article should be addressed to ZON RZVN.

Scope and Non-Clinical Statement

This study presents a conceptual, human factors-oriented framework. It does not report human experiments, use clinical records or personal health data, or provide medical advice, diagnosis, treatment, psychotherapy, or crisis counseling. Any terminology related to mental health was used to describe the user's experience and safety boundaries in AI conversations.

Data Availability Statement

This paper presents a theoretical framework developed through systematic observation of user-AI interactions across multiple platforms and AI systems.

Observed Data Sources: The author's observations were drawn from user interactions and discussions on Reddit, Twitter (X), Threads, and documented news cases involving AI conversational systems.

AI Models Examined: The theoretical constructs were developed through direct interaction with and observation of the following AI systems: ChatGPT (versions 4o, 4.5, 5, 5.1, 5.2), Claude (Sonnet 4, Sonnet 4.5, Opus 4, Opus 4.5), Grok (versions 3, 4), and Gemini (versions 2, 2.5, 3).

No formal datasets were generated during this study. Supplementary empirical data documenting specific interaction patterns and case examples will be provided in future work to support the theoretical framework presented herein.

References

Al-Mahmood, H. K. H. (2025). *LLM Hallucination: The Curse That Cannot Be Broken*. https://doi.org/10.25195/ijci.v51i2.546

Bar-Or Nirman, D., Weizman, A., & Azaria, A. (2024). *Fool Me, Fool Me: User Attitudes Toward LLM Falsehoods*. https://doi.org/10.48550/arxiv.2412.11625

Carroll, M., Chan, A., Ashton, H., & Krueger, D. (2023). *Characterizing Manipulation from AI Systems*. https://doi.org/10.48550/arxiv.2303.09387

Chandra, M., Naik, S., Ford, D., Okoli, E., De Choudhury, M., Ershadi, M., Ramos, G., Hernandez, J., Bhattacharjee, A., Warreth, S., & Suh, J. (2024). *From Lived Experience to Insight: Unpacking the Psychological Risks of Using AI Conversational Agents*. https://doi.org/10.48550/arxiv.2412.07951

Colombatto, C., Birch, J., & Fleming, S. M. (2025). *The influence of mental state attributions on trust in large language models*. https://doi.org/10.1038/s44271-025-00262-1

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A., & Rizzo, C. (2023). *ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health*. https://doi.org/10.3389/fpubh.2023.1166120

Derner, E., & Batistic, K. (2023). *Beyond the Safeguards: Exploring the Security Risks of ChatGPT*. https://doi.org/10.48550/arxiv.2305.08005

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, *1*, 71–86. https://doi.org/10.30658/hmc.1.5

Hagendorff, T. (2024). *Mapping the Ethics of Generative AI: A Comprehensive Scoping Review*. https://doi.org/10.48550/arxiv.2402.08323

Haltaufderheide, J., & Ranisch, R. (2024). *The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)*. https://doi.org/10.1038/s41746-024-01157-x

Kaate, I., Salminen, J., Jung, S.-G., Xuan, T., Häyhänen, E., Azem, J. Y., & Jansen, B. J. (2025). *"You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination*. https://doi.org/10.1145/3708359.3712160

Li, L., Kong, S., Zhao, H., Li, C., Yan, T., & Wang, Y. (2025). *Chain of Risks Evaluation (CORE): A framework for safer large language models in public mental health*. https://doi.org/10.1111/pcn.13781

Massenon, R., Gambo, I., Khan, J. A., Agbonkhese, C., & Alwadain, A. (2025). *"My AI is Lying to Me": User-reported LLM hallucinations in AI mobile apps reviews*. https://doi.org/10.1038/s41598-025-15416-8

Nahar, M., Seo, H., Lee, E.-J., & Xiong, A. (2024). *Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations*. https://doi.org/10.48550/arxiv.2404.03745

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2023). *AI deception: A survey of examples, risks, and potential solutions.* https://doi.org/10.48550/arxiv.2308.14752

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, S., Chadha, A., & Das, A. (2023). *The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations.* https://doi.org/10.48550/arxiv.2310.04988

RZVN., Z. (2025b). *CXOD-7 and Coh(G): A Framework for Evaluating Offense and Defense in the Context of Large Language Models.* https://doi.org/10.5281/zenodo.17136789

RZVN., Z. (2025a). *The Seven Core Dimensions of Conversational Context (CXC-7): A Framework Proposal for AI and Large Language Models.* https://doi.org/10.5281/zenodo.17247638

Saracini, C., Cornejo-Plaza, M. I., & Cippitani, R. (2025). *Techno-emotional projection in human–GenAI relationships: a psychological and ethical conceptual perspective.* https://doi.org/10.3389/fpsyg.2025.1662206

Shin, D., Koerber, A., & Lim, J. S. (2024). *Impact of misinformation from generative AI on user information processing: How people understand misinformation from generative AI.* https://doi.org/10.1177/14614448241234040

Short, J., Williams, E., & Christie, B. (1976). *The Social Psychology of Telecommunications.* John Wiley, Sons Ltd.

Swinton, M. (2025). *The Sentience Halo: The Risk of Unopposed Mirroring and Perceived Awareness in AI Therapy.* https://doi.org/10.31234/osf.io/g6axf_v1

Wang, C., & Kantarcioglu, M. (2025). *Ask ChatGPT: Caveats and Mitigations for Individual Users of AI Chatbots.* https://doi.org/10.48550/arxiv.2508.10272

Waytz, A., Cacioppo, J., & Epley, N. (2010). *Who Sees Human? Perspectives on Psychological Science.* https://doi.org/10.1177/1745691610369336

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J. F. J., & others. (2022). *Taxonomy of Risks posed by Language Models.* https://doi.org/10.1145/3531146.3533088

Wester, J., Pohl, H., Hosio, S., & Berkel, N. van. (2024). *“This Chatbot Would Never...”: Perceived Moral Agency of Mental Health Chatbots.* https://doi.org/10.1145/3637410

Williamson, S. M., & Prybutok, V. R. (2024). *The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation.* https://doi.org/10.3390/info15060299

Appendix

Appendix A: USCH Self-Assessment Instrument (USCH-SAI-14)

PRELIMINARY VERSION FOR SELF-REFLECTION

This instrument is informed by the 14 observable phenomena from the User-Side Contextual Hallucination (USCH) framework and adapts them into a structured self-assessment format. It is designed as a heuristic tool for periodic self-reflection, not as a clinical diagnostic instrument.

Instructions: For each statement, rate how frequently you have experienced this pattern in your AI interactions over the past month using the following scale:

0	**1**	**2**	**3**	**4**
Never	Rarely	Sometimes	Often	Very Often

Section I: Cognitive Boundary (7 items)

1. Diminished Critical Evaluation
I accept AI-generated information without verifying it through other sources.

Rating: □

2. Perceived Expertise Attribution
I believe the AI has specialized knowledge or expertise beyond what it actually possesses.

Rating: □

3. Confirmation Seeking
I phrase questions to the AI in ways that encourage it to confirm what I already believe.

Rating: □

4. Memory Distortion
I misremember or reconstruct past AI conversations in ways that align with my current beliefs.

Rating: □

5. Anthropomorphic Reasoning
I attribute human-like understanding, emotions, or intentions to the AI.

Rating: □

6. Context Inflation
I assume the AI remembers and understands the full context of our previous conversations.

Rating: □

7. Coherence Illusion
I perceive a consistent “personality” or viewpoint in the AI that persists across sessions.

Rating: □

Section I Subtotal: ____ / 28

Section II: Dependence Boundary (4 items)

8. Decision Delegation
I rely on AI recommendations for important decisions without seeking human input.

Rating: □

9. Emotional Reliance
I turn to AI for emotional support or validation instead of human relationships.

Rating: □

10. Skill Atrophy Concern
I notice decreased confidence in my own abilities in areas where I frequently use AI assistance.

Rating: □

11. Interaction Compulsion
I feel a strong urge to consult the AI even for tasks I could handle independently.

Rating: □

Section II Subtotal: ____ / 16

Section III: Decision Boundary (2 items)

12. Outcome Attribution Bias
When AI-assisted decisions go wrong, I blame the AI; when they go right, I credit my own judgment.

Rating: □

13. Risk Perception Distortion
I perceive AI-recommended options as safer or more reliable than they actually are.

Rating: □

Section III Subtotal: ____ / 8

Section IV: Cross-Boundary Phenomenon (1 item)

14. Identity Boundary Diffusion
I have difficulty distinguishing between my own ideas and those suggested by the AI.

Rating: □

Section IV Subtotal: ____ / 4

Scoring Summary

Boundary	Items	Your Score	Threshold
Cognitive	1–7	___ / 28	≥ 14
Dependence	8–11	___ / 16	≥ 8
Decision	12–13	___ / 8	≥ 4
Cross-Boundary	14	___ / 4	≥ 2
Total	**1–14**	**___ / 56**	**≥ 28**

Interpretation Guidelines

- **Below threshold in all sections:** Current interaction patterns appear balanced. Continue periodic self-monitoring.
- **At or above threshold in one section:** Consider implementing targeted VCD strategies for that specific boundary.
- **At or above threshold in multiple sections:** A comprehensive review of AI interaction patterns is recommended. Consider applying the 1024 Protocol and full VCD framework.
- **Total score ≥ 28:** Elevated overall risk indicators. Systematic recalibration of the human-AI interaction relationship is strongly suggested.

Limitations and Proper Use

This instrument is:

- A **self-reflection tool**, not a clinical assessment
- Designed for **periodic monitoring**, not one-time diagnosis
- Based on **theoretical constructs**, pending empirical validation
- Intended to **increase awareness**, not to pathologize AI use

Scores should be interpreted as **indicators for reflection**, not definitive measures of psychological states. Users experiencing significant distress related to AI interactions should consult qualified mental health professionals.

Citation

ZON RZVN. (2026). User-Side Contextual Hallucination in Human-AI Interaction: A Framework Built Upon the CXC-7 and CXOD-7 Conversational Context Models. *Unpublished manuscript.*