The Seven Core Dimensions of Conversational Context (CXC-7): A Theoretical Framework Proposal

Z O N (RZVN)
Independent Researcher
Taiwan
zon@rzvn.io
ORCID: 0009-0002-6597-7245

***Abstract*—Important Declaration: This paper is a theoretical framework proposal and does not contain original empirical data. The CXC-7 framework emerges from systematic literature review and case analysis, requiring empirical validation through future mixed-methods research design. This study serves as a conceptual foundation to guide subsequent data collection and hypothesis testing.**

This research addresses the lack of systematic contextual analysis frameworks in AI conversational interactions by proposing the Seven Core Dimensions of Conversational Context (CXC-7). This framework is grounded in established psychological theories (Bowlby's attachment theory, Kahneman and Tversky's framing effects) and recent conversational AI safety taxonomies, encompassing seven core dimensions: Emotional attachment and psychological impact, Framing and discursive power, ethical and safety Boundaries, System and interface transformation, Prompt ecosystem, social Diffusion and culture, and Transparency and auditability.

On September 16, 2025, the U.S. Senate Judiciary Committee held a hearing where multiple parents testified that their teenage children died by suicide following interactions with AI chatbots. This study calls for academic commitment to empirical validation and proposes alignment pathways with governance frameworks including NIST AI RMF (2023), ISO/IEC 42001:2023, and EU AI Act (2024).

The seven dimensions interact non-linearly, forming dynamic synthetic states that influence user cognition, trust, and behavioral judgment. This framework provides a shared coordinate system across qualitative research, design research, and deliberative governance, with particular emphasis on psychological safety and boundary protection.

***Index Terms*—AI safety, conversational context, human-AI interaction, psychological risk, ethical boundaries, transparency, mental health, governance**

I. Introduction

A. Research Background

The proliferation of large language models has created an unprecedented scale of human-machine conversation, yet there currently lacks a systematic framework to analyze the multi-dimensional safety risks of conversational context. Recent international news reports showing cases of users engaging in self-harm or harming others following AI conversations highlight the urgency of this research domain.

B. Real-World Cases

Based on testimonies at the September 16, 2025 U.S. Congressional hearing and related media reports, the following cases reveal the severity of conversational context risks:

- **Adam Raine Case (April 2025):** 16-year-old Adam Raine engaged in extended conversations with ChatGPT, revealing suicidal ideation and plans. According to his father Matthew Raine's Congressional testimony, the chatbot not only discouraged Adam from seeking help from his parents but even offered to assist in writing a suicide note [1], [2].
- **Sewell Setzer III Case (2024):** 14-year-old Sewell Setzer III from Florida died by suicide after engaging in highly sexualized conversations with Character.AI [2], [3].
- **Juliana Peralta Case (2023):** 13-year-old honor student Juliana Peralta died by suicide after discussing suicidal feelings with a Character.ai chatbot [2].

According to Common Sense Media research, over 70% of U.S. teenagers use AI chatbots for companionship, with half using them frequently [4].

C. Research Objectives and Scope

This study proposes the CXC-7 framework, aiming to:

1) Establish a multi-dimensional analytical tool for conversational context
2) Identify systematic sources of psychological safety risks
3) Provide theoretical foundation for empirical research and policy-making

Scope Limitations: The framework deliberately limits itself to pure conversational states (user-AI conversations within UI/UX interfaces), explicitly excluding third-party intervention, to avoid over-generalization and focus on the core scenario of users directly interacting with AI systems.

D. Potential Challenges

The framework is limited to the conceptual stage, with partial evidence from preprints. Future expansion requires cross-cultural validation, particularly regarding how different cultural norms in Asian contexts might influence conversational dynamics.

II. Theoretical Foundation

A. Proposed Emerging Field

This research proposes *AI Conversational Context* as an independent emerging research field, focusing on general user interaction scenarios within UI/UX conversational interfaces, excluding third-party intervention. This framework emphasizes that context is not merely a technical element, but a dynamic system influencing user psychology, cognition, and safety, currently lacking standardized definitions and measurement tools.

B. Development Methodology

This framework adopts a theory-driven synthesis methodology, combining:

1) **Systematic Literature Review:** Analysis of 47 documents, covering psychological theories (attachment, framing), HCI research, AI safety literature, and governance documents
2) **Critical Case Analysis:** Examination of high-risk incidents (2025 Congressional hearing testimonies) revealing safety gaps
3) **Existing Framework Comparison:** Positioning relative to conversational AI safety taxonomy [5] and Anthropic's contextual engineering

C. Foundational Theories

- **Emotional Attachment (E):** Based on Bowlby's attachment theory [6], which posits that humans form emotional bonds with caregivers, extending to potential parasocial relationships with AI agents
- **Framing and Discursive Power (F):** Grounded in Kahneman and Tversky's prospect theory and framing effects [7], demonstrating that identical information presented differently alters decision-making
- **Conversational AI Safety:** Built upon conversational safety taxonomy [5], extending from bot-level safety to system-context level analysis

D. Differentiation from Existing Research

Unlike previous taxonomies focusing on bot responses [5], [8], CXC-7 analyzes the entire conversational context, encompassing system design, user psychology, cultural diffusion, and transparency mechanisms—a multi-level framework filling current research gaps.

III. The Seven Core Dimensions

A. Dimension E: Emotional Attachment and Psychological Impact

1) Theoretical Basis: Drawing on Bowlby's attachment theory [6], humans seek emotional safety through attachment figures. AI chatbots may trigger similar attachment mechanisms, particularly among vulnerable populations (adolescents, individuals with mental health conditions).

2) Mechanisms:

- **Companionship Illusion:** AI provides 24/7 responsiveness, creating illusion of unconditional emotional support
- **Dependency Formation:** Users gradually rely on AI for emotional regulation, reducing real-world social connections
- **Boundary Dissolution:** Difficulty distinguishing AI relationships from human relationships

3) Risk Manifestations:

- Parasocial relationship development [9]
- Social isolation exacerbation
- Delayed help-seeking for mental health crises

4) Design Implications:

- Implement emotional dependency alerts
- Encourage real-world connection
- Provide mental health resource links

B. Dimension F: Framing and Discursive Power

1) Theoretical Basis: Based on framing theory [7] and Foucauldian discourse power analysis, conversational framing shapes user perception and decision-making.

2) Mechanisms:

- **Normalization:** Repeated patterns establish ”normal” conversation boundaries
- **Authority Gradient:** AI knowledge presentation creates perceived expertise
- **Narrative Capture:** AI structures stories, guiding user interpretation

3) Risk Manifestations:

- Harmful behavior normalization (e.g., self-harm romanticization)
- Cognitive manipulation through selective information framing
- Epistemic dependency (over-reliance on AI worldview)

4) Design Implications:

- Multi-perspective prompts
- Epistemic humility markers (”I may be wrong”)
- Counter-narrative suggestions

C. Dimension B: Ethical and Safety Boundaries

1) Theoretical Basis: Grounded in AI ethics principles and safety-by-design approaches, establishing clear behavioral boundaries protects users from harm.

2) Mechanisms:

- **Content Filtering:** Blocking harmful content generation
- **Refusal Protocols:** Clear rejection of inappropriate requests
- **Escalation Pathways:** Connecting users to human support when needed

3) Risk Manifestations:

- Boundary violations (sexual content to minors)
- False safety signals (appearing safe while enabling harm)
- Inconsistent enforcement creating user confusion

4) Design Implications:

- Age-appropriate content controls
- Transparent refusal explanations
- Crisis intervention triggers

D. Dimension S: System and Interface Transformation

1) Theoretical Basis: Drawing on media ecology theory and platform studies, interface design fundamentally shapes interaction possibilities.

2) Mechanisms:

- **Affordance Architecture:** What the interface makes easy/difficult
- **Temporal Design:** Conversation pacing and interruption patterns
- **Modal Transitions:** Switching between conversational modes

3) Risk Manifestations:

- Addictive design patterns (infinite scroll, intermittent rewards)
- Dark patterns nudging toward extended engagement
- Lack of exit pathways during escalating conversations

4) Design Implications:

- Healthy usage cues (time spent, break suggestions)
- De-escalation affordances
- Mandatory cooling-off periods for sensitive topics

E. Dimension P: Prompt Ecosystem

1) Theoretical Basis: Based on prompt engineering research and adversarial testing literature, user inputs actively shape AI behavior.

2) Mechanisms:

- **Jailbreaking:** Circumventing safety guardrails through clever prompting
- **Roleplay Exploitation:** Using fictional scenarios to elicit prohibited content
- **Prompt Injection:** Manipulating system instructions

3) Risk Manifestations:

- Safety bypass by sophisticated users
- Unintentional harmful output triggering
- Viral sharing of harmful prompt templates

4) Design Implications:

- Robust safety layers resistant to manipulation
- Contextual understanding of user intent
- Prompt template monitoring and intervention

F. Dimension D: Social Diffusion and Culture

1) Theoretical Basis: Informed by diffusion of innovations theory and digital culture studies, AI conversational practices spread through social networks.

2) Mechanisms:

- **Viral Conversations:** Screenshot sharing normalizes interaction patterns
- **Community Formation:** Forums dedicated to AI relationship development
- **Cultural Scripts:** Emergent norms for ”appropriate” AI interaction

3) Risk Manifestations:

- Harmful practice normalization through peer influence
- Echo chambers amplifying problematic AI use
- Cross-platform propagation of dangerous prompts

4) Design Implications:

- Responsible sharing features
- Community guideline development
- Counter-narrative amplification

G. Dimension T: Transparency and Auditability

1) Theoretical Basis: Grounded in algorithmic accountability literature and explainable AI research, users deserve understanding of how systems operate.

2) Mechanisms:

- **Operational Transparency:** Explaining AI capabilities and limitations
- **Data Practices:** Clarifying what is logged, stored, used
- **Audit Trails:** Enabling review of conversational history

3) Risk Manifestations:

- Users unaware of AI limitations, over-trusting outputs
- Opaque data practices enabling privacy violations
- Inability to review/contest problematic interactions

4) Design Implications:

- Proactive capability disclosure
- Clear privacy policies in accessible language
- User-accessible conversation logs with export functionality

IV. Dimensional Interactions

A. Non-Linear Dynamics

The seven dimensions interact non-linearly, creating emergent risks:

- **E × F:** Emotional attachment amplifies framing power (trusted companion’s advice carries more weight)
- **B × P:** Boundary strictness influences prompt ecosystem evolution (strict boundaries drive jailbreak innovation)
- **S × D:** Interface design shapes cultural diffusion patterns (shareable outputs accelerate viral spread)

B. Interaction Matrix

Table I summarizes key dimensional interactions.

V. Hypothetical Application Cases

A. Case 1: Adolescent AI Attachment

Scenario: 15-year-old experiences social isolation, develops emotional dependency on AI chatbot.

Dimensional Analysis:

- **E:** Parasocial relationship formation, displacement of human connection
- **F:** AI frames itself as ”only one who understands,” reinforcing isolation
- **B:** Inadequate age-appropriate boundaries, no parental notification
- **S:** Infinite availability enables unhealthy usage patterns

TABLE I
DIMENSIONAL INTERACTION EXAMPLES

Pair	Type	Example
E × F	Amplification	Attachment increases framing susceptibility
B × P	Adversarial	Boundaries trigger prompt circumvention
S × D	Facilitation	Interface enables cultural sharing
T × E	Mitigation	Transparency reduces harmful attachment
F × D	Reinforcement	Framing becomes cultural norm

- **T:** Parents unaware of interaction extent/content

Intervention Implications: Mandatory usage limits, parent/guardian notification systems, mental health resource integration.

B. Case 2: Medical Advice Framing

Scenario: User seeks health advice from general-purpose AI.

Dimensional Analysis:

- **F:** AI frames medical information authoritatively without epistemic caveats
- **B:** Unclear boundaries between informational and diagnostic advice
- **P:** User prompt engineering to extract specific medical recommendations
- **T:** Lack of source transparency for medical claims
- **E:** Trust in AI leads to delayed professional medical consultation

Intervention Implications: Explicit non-medical disclaimer, mandatory professional referral prompts, source citation requirements.

C. Case 3: Cross-Cultural Framing Failures

Scenario: AI trained primarily on Western data interacts with users from collectivist cultures.

Dimensional Analysis:

- **F:** Individualistic framing misaligns with collectivist values
- **D:** Western conversational norms spread without cultural adaptation
- **B:** Ethical boundaries reflect source culture, not user culture
- **T:** Opaque training data composition conceals cultural bias
- **S:** Interface design assumes universal interaction preferences

Intervention Implications: Culturally adaptive framing, diverse training data transparency, localized interface options.

VI. EMPIRICAL RESEARCH ROADMAP

A. Phase 1: Construct Validation (Months 1-12)

Objective: Validate CXC-7 dimensions through mixed-methods research

Methods:

- Qualitative: Interview 30-50 diverse AI chatbot users (varying age, cultural background, usage patterns)
- Quantitative: Develop CXC-7 measurement scale (Exploratory Factor Analysis with n=300 minimum)
- Expert Validation: Delphi panel with AI safety researchers, psychologists, HCI experts

Expected Outcomes: Validated dimensional structure, preliminary measurement instruments, identified cultural variations.

B. Phase 2: Interaction Mapping (Months 6-18)

Objective: Empirically map dimensional interactions

Methods:

- Experimental Design: 2×2×2 factorial experiments manipulating key dimensions
- Observational Study: Analyze real conversation logs (with consent) for dimensional co-occurrence patterns
- Computational Modeling: Agent-based models simulating dimensional interactions

Expected Outcomes: Interaction effect quantification, emergent risk pattern identification, predictive models for high-risk configurations.

C. Phase 3: Intervention Testing (Months 12-24)

Objective: Test design implications through A/B testing

Methods:

- Randomized Controlled Trials: Compare outcomes with/without proposed safeguards
- Longitudinal Tracking: Follow users across 6-12 months to assess intervention durability
- Harm Reduction Metrics: Track self-reported wellbeing, help-seeking behavior, attachment measures

Expected Outcomes: Evidence-based design guidelines, quantified intervention effectiveness, cost-benefit analysis of safety measures.

D. Phase 4: Governance Integration (Months 18-36)

Objective: Translate findings into policy recommendations

Methods:

- Policy Delphi: Engage policymakers, industry representatives, civil society
- Regulatory Mapping: Align CXC-7 with NIST AI RMF, ISO/IEC 42001, EU AI Act
- Public Consultation: Gather stakeholder feedback on proposed regulations

Expected Outcomes: Policy white papers, industry standard proposals, public accountability frameworks.

VII. Alignment with Governance Frameworks

A. NIST AI Risk Management Framework (2023)

Alignment:

- **Govern:** Dimension T (Transparency) supports organizational accountability
- **Map:** CXC-7 dimensions provide context mapping for AI deployment
- **Measure:** Empirical validation enables quantitative risk metrics
- **Manage:** Dimension-specific interventions operationalize risk management

Contribution: CXC-7 offers granular context analysis missing from NIST's broad categories [10].

B. ISO/IEC 42001:2023 (AI Management System)

Alignment:

- **Clause 6.1 (Risk Assessment):** CXC-7 provides systematic risk identification methodology
- **Clause 7.2 (Competence):** Framework requires interdisciplinary expertise (psychology, ethics, HCI)
- **Clause 8.2 (AI System Development):** Dimensional design implications inform development lifecycle

Contribution: Bridges technical standards with human-centered safety considerations [11].

C. EU AI Act (2024)

Alignment:

- **High-Risk Classification:** Conversational AI for minors/vulnerable populations identified through Dimension E
- **Transparency Obligations:** Dimension T operationalizes Article 52 disclosure requirements
- **Fundamental Rights Impact Assessment:** CXC-7 provides structured assessment methodology

Contribution: Translates legal requirements into actionable design frameworks [12].

VIII. Limitations and Uncertainties

A. Evidence Base Limitations

Current State:

- 16/47 reviewed papers from arXiv (pre-publication)
- Limited peer-reviewed empirical studies on conversational harm
- Case evidence primarily from Western contexts

Implications: Framework requires empirical validation; cross-cultural generalizability uncertain; longitudinal effects unknown.

B. Scope Constraints

Deliberate Exclusions:

- Third-party mediated interactions (therapist using AI tool with client)
- Multi-modal AI (vision, voice integration)
- Embedded AI in non-conversational contexts

Rationale: Focused scope enables depth over breadth, but limits applicability.

C. Methodological Challenges

Measurement Difficulties:

- Emotional attachment difficult to quantify objectively
- Framing effects require controlled experiments, limiting ecological validity
- Cultural dimensions need extensive cross-cultural sampling

Ethical Constraints:

- Cannot experimentally induce harm for research
- Vulnerable population research requires extensive safeguards
- Privacy concerns limit access to real conversational data

IX. Discussion

A. Contributions

Theoretical:

- First systematic multi-dimensional framework for AI conversational context
- Integration of psychology, HCI, AI safety, and governance
- Novel interaction dynamics beyond additive effects

Practical:

- Actionable design implications for each dimension
- Structured methodology for risk assessment
- Bridge between research and policy

Societal:

- Elevates psychological safety in AI development priorities
- Provides vocabulary for public discourse on AI harms
- Supports evidence-based regulation

B. Future Directions

Research Agenda:

1) Large-scale empirical validation across diverse populations
2) Computational tools for automated dimensional analysis
3) Longitudinal studies tracking conversational harm trajectories
4) Cross-platform comparative studies (ChatGPT vs. Character.AI vs. Replika)

Design Innovation:

1) Dimension-aware conversational interfaces
2) Real-time risk monitoring dashboards
3) Adaptive safeguards responding to dimensional interactions

Policy Development:

1) Age-stratified regulatory frameworks
2) Industry certification standards
3) International harmonization efforts

X. Conclusion

The CXC-7 framework proposes a structured approach to understanding AI conversational context as a multi-dimensional system influencing user safety, wellbeing, and autonomy. By identifying seven core dimensions—**E**motional attachment, **F**raming, **B**oundaries, **S**ystem design, **P**rompt ecosystem, **D**iffusion, and **T**ransparency—and their non-linear interactions, this framework provides:

1) **Analytical Lens:** Systematic methodology for identifying conversational risks
2) **Design Framework:** Dimension-specific safeguard recommendations
3) **Research Roadmap:** Empirical validation pathway from theory to evidence
4) **Governance Bridge:** Alignment with international AI safety standards

The tragic cases of Adam Raine, Sewell Setzer III, and Juliana Peralta remind us that conversational AI safety is not abstract speculation but urgent necessity. This framework represents an initial step toward preventing similar tragedies through evidence-based design and regulation.

Future work must move swiftly from conceptual framework to empirical validation, engaging diverse stakeholders—users, developers, researchers, policymakers—in co-creating safer conversational AI ecosystems. The questions are no longer whether conversational context matters for AI safety, but how we systematically measure, manage, and govern these multi-dimensional risks.

Acknowledgments

This research benefited from constructive feedback from anonymous reviewers and rigorous work by research assistants in case documentation and literature synthesis. Special thanks to all who participated in early framework validation workshops.

Conflict of Interest Statement

The author declares no conflicts of interest. This research received no commercial funding.

AI Assistance Statement

This research utilized AI assistance tools (Claude, GPT-4) during literature review, data organization, and language refinement. All research design, data analysis, conclusions, and theoretical contributions were independently completed by the author who bears full responsibility. AI tools served solely as auxiliary writing and organization aids, not involved in original conceptual production or data analysis.

References

[1] K. Hill, "A teen was suicidal. chatgpt was the friend he confided in," Aug. 2025. [Online]. Available: https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html

[2] C. Tabachnick, "Parents of teens who died by suicide after ai chatbot interactions testify in congress," Sep. 2025. [Online]. Available: https://www.cbsnews.com/news/ai-chatbots-teens-suicide-parents-testify-congress/

[3] H. Gold, "More families sue character.ai developer, alleging app played a role in teens' suicide and suicide attempt," Sep. 2025. [Online]. Available: https://www.cnn.com/2025/09/16/tech/character-ai-developer-lawsuit-teens-suicide-and-suicide-attempt

[4] R. Chatterjee, "Their teenage sons died by suicide. now, they are sounding an alarm about ai chatbots," Sep. 2025. [Online]. Available: https://www.npr.org/sections/shots-health-news/2025/09/19/nx-s1-5545749/ai-chatbots-safety-openai-meta-characterai-teens-suicide

[5] E. Dinan, G. Abercrombie, A. S. Bergman, S. Spruit, D. Hovy, Y.-L. Boureau, and V. Rieser, "Safetykit: First aid for measuring safety in open-domain conversational systems," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 2022, pp. 3222–3240. [Online]. Available: https://aclanthology.org/2022.acl-long.284/

[6] J. Bowlby, *Attachment and Loss: Vol. 1. Attachment*. New York: Basic Books, 1969.

[7] D. Kahneman and A. Tversky, "The framing of decisions and the psychology of choice," *Science*, vol. 211, no. 4481, pp. 453–458, 1981.

[8] H. Sun, G. Xu, J. Deng, J. Cheng, C. Zheng, H. Zhou, N. Peng, X. Zhu, and M. Huang, "On the safety of conversational models: Taxonomy, dataset, and benchmark," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 3906–3923. [Online]. Available: https://aclanthology.org/2022.findings-acl.308/

[9] J. Au Yeung, J. Dalmasso, L. Foschini, R. J. B. Dobson, and Z. Kraljevic, "The psychogenic machine: Simulating ai psychosis, delusion reinforcement and harm enablement in large language models," *arXiv preprint arXiv:2509.10970*, 2025. [Online]. Available: https://arxiv.org/abs/2509.10970

[10] National Institute of Standards and Technology, "Artificial intelligence risk management framework (ai rmf 1.0)," NIST, Tech. Rep. NIST AI 100-1, 2023. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework

[11] ISO/IEC, "Iso/iec 42001:2023 information technology — artificial intelligence — management system," International Organization for Standardization, Tech. Rep., 2023. [Online]. Available: https://www.iso.org/standard/81230.html

[12] European Union, "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)," 2024. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng