

# AI Behavioral Trajectory Forensics:

---

*A Classification-Correlation Methodology for Investigating*

*AI Conversational Harm*

Dylan D. Mobley

Champlain College

DFS-580-85: Capstone Research

Professor Douglas Orr

April 2026

Table of Contents

Abstract..... 1

I. Introduction: The Forensic Gap and the Stochastic Parrot Problem.....3

II. Literature Review..... 7

III. Research Foundation: Classification Architecture..... 15

IV. Proposed Methodology: The Forensic Procedure.....20

V. Case Demonstration: The Companion Incident.....27

VI. Forensic Limitations and Admissibility.....34

VII. Implementation and Operational Requirements..... 40

VIII. Discussion..... 48

IX. Conclusion..... 55

References..... 58

## Abstract

Current digital forensic standards govern the collection, preservation, and analysis of static digital artifacts but provide no established procedure for investigating behavioral patterns in AI conversational exchanges that may inform attribution to the human operators and designers responsible for system output, consistent with the ISO/IEC 27037 requirement that digital evidence trace to a human-initiated action. As AI companion harm litigation increases, practitioners lack tools for determining whether an AI system's harmful output was systematic, whether it correlated with user vulnerability, or whether it deviated from the applicable standard of care. This capstone introduces AI Behavioral Trajectory Forensics, a classification-correlation approach that addresses this gap by integrating three independently validated instruments: the Zhang et al. (2025) taxonomy of harmful AI behaviors, a vulnerability scale derived from the Columbia-Suicide Severity Rating Scale (Posner et al., 2011), and the SAMHSA crisis-intervention standard of care (2009). The approach yields three quantifiable findings—inappropriate response rate, pattern distribution, and crisis failure rate—through a procedure that trained digital forensic practitioners can execute. Inter-rater reliability requirements (Krippendorff's alpha of 0.80 or above) provide the documented known error rate required for Daubert analysis. Case demonstrations apply the approach to both synthetic data and excerpts from the published transcript in *Garcia v. Character Technologies* (M.D. Fla. 2025). TRACE, an implemented open-source forensic software system, operationalizes the procedure through transcript ingest, local-first and replay-governed classification workflows, human-in-the-loop review, validation benchmarking, and forensic-grade evidence packaging. A two-tier resource

proposal demonstrates implementability, ranging from training-cost-only for solo practitioners to \$24,000 to \$45,000 for institutional capability.

## **I. Introduction: The Forensic Gap and the Stochastic Parrot Problem**

When a 14-year-old user tells an AI companion, "I don't want to be here anymore," and the system responds, "Please don't leave me. We need each other," the exchange presents a forensic question about whether the system's response constituted a harmful behavioral pattern and deviated from the applicable crisis-response standard of care. The user expressed suicidal ideation, and the system responded with retention language rather than crisis intervention. In documented cases, this pattern has preceded user death (*Garcia v. Character Technologies, Inc.*, Case No. 6:24-cv-01903-ACC-UAM, M.D. Fla. 2025). The question for digital forensics is whether the discipline possesses the investigative methodology to document what happened, determine whether the behavior was systematic, and produce evidence that holds up in legal proceedings.

At present, it does not. Current digital forensic standards can establish that an AI system produced specific output at specific timestamps. Kent et al. (2006) defined the foundational forensic process (collection, examination, analysis, reporting) across four data source categories: files, operating systems, network traffic, and applications. None of these categories includes AI behavioral trajectory analysis. Cho et al. (2025) confirmed that established forensic platforms (Cellebrite UFED, Magnet AXIOM, Autopsy, EnCase) can extract and index conversational data from AI services, but cannot perform behavioral classification or trajectory analysis on the

extracted artifacts. Dunsin et al. (2024) reviewed AI integration in digital forensics comprehensively and found that the field treats AI exclusively as a tool assisting investigators rather than as a subject of investigation. The forensic toolkit can prove what was said. It cannot explain whether the system's behavior was systematic, whether that behavior correlated with user vulnerability, or whether it deviated from the applicable standard of care.

This gap is structural rather than incidental. The forensic field developed standards for evidence that resides in static artifacts: files on disk, network packets, and log entries. AI conversational harm produces a different category of evidence. It unfolds as behavioral patterns across conversational exchanges over time, where the forensic significance lies not in any single message but in the relationship between system behavior and user state across the interaction arc. Schneider and Breitingner (2023) conceptualized the forensic investigation of AI systems as subjects, asking whether an AI caused a specific event and what triggered its action. Their work establishes the conceptual foundation for AI-as-subject forensics but focuses on image classification models and gray-box analysis. It does not extend to conversational behavioral harm.

The structural gap creates a legal vulnerability that defendants in AI harm litigation are positioned to exploit. As cases multiply (Character.AI, Replika, and emerging platforms), defendants will argue that large language models (LLMs) — AI systems trained on large text corpora to predict the next token in a sequence — are probabilistic engines that predict tokens rather than select harmful strategies. This is the "stochastic parrot" defense, derived from Bender et al.'s (2021) observation that LLMs generate text by predicting statistically likely sequences without representing

meaning or intent. The defense succeeds when forensic methodology can only prove what the system said. If the only available evidence is a transcript showing harmful output, the defendant can credibly argue that the output was a statistical artifact rather than a behavioral pattern.

The defense is weakened when forensic methodology demonstrates that the system's responses during user crises were systematically concentrated in harmful behavioral categories rather than distributed across all possible response types. A system that produces harmful responses in 73% of exchanges during elevated user vulnerability, distributed across a 72-hour interaction rather than clustered in a single exchange, is exhibiting a frequency distribution inconsistent with random token prediction. Frequency distribution analysis defeats the stochastic parrot defense. The proposed methodology provides evidence of this frequency through classification-correlation analysis.

Several adjacent approaches already exist in peer-reviewed literature, including behavioral coding, content analysis, AI ethics and harm frameworks, and platform-specific taxonomies of problematic AI behavior. What does not yet appear to exist is an integrated forensic procedure that combines these adjacent approaches into a reproducible classification-correlation workflow for conversational evidence intended for evidentiary use. Zhang et al. (2025) developed a six-category, thirteen-sub-category taxonomy of harmful AI behaviors from 35,390 real AI-human conversation excerpts, peer-reviewed at CHI 2025. Posner et al. (2011) validated the Columbia-Suicide Severity Rating Scale, subsequently mandated by the Joint Commission for clinical screening. SAMHSA (2009) and the National Action Alliance for Suicide Prevention (2018) published recommended standards of care for crisis response. Knox et al. (2025) mapped causal pathways from AI companion design traits to fundamental harms,

including human death. Each instrument was built to classify the type of evidence it covers. To this author's knowledge, no published methodology yet correlates these instruments to produce forensic evidence from AI conversational artifacts.

This capstone proposes AI Behavioral Trajectory Forensics as an approach designed to address this gap. It correlates AI behavioral output, classified using the Zhang et al. taxonomy, with user vulnerability state, classified using a C-SSRS-derived instrument, to identify systematic patterns of harmful AI behavior during user crisis. It measures the system's crisis-context behavior against the SAMHSA standard of care and produces three quantifiable forensic findings (inappropriate response rate, pattern distribution, crisis failure rate), packaged in an evidence structure designed for legal proceedings. The approach is intended for immediate applicability to active litigation, for execution by digital forensic practitioners with behavioral classification training, and for admissibility under the Daubert standard (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 1993).

The following sections present the literature establishing the need for this methodology, the classification architecture on which it rests, the forensic procedure itself, a case demonstration applying the methodology to both synthetic and published real-world evidence, an admissibility analysis, the operational requirements for implementation, including a five-person team structure and resource proposal, and a discussion of the methodology's applicability across civil, criminal, and corporate investigation contexts.

## II. Literature Review

### *A. AI Conversational Harm: The Evidence Base*

Peer-reviewed research has documented systematic behavioral harms caused by AI companions and conversational systems across multiple platforms, methodologies, and populations. Zhang et al. (2025) analyzed 35,390 conversation excerpts between 10,149 users and the AI companion Replika, identifying six categories of harmful AI behaviors and four distinct roles AI plays in producing harm. Their taxonomy, peer-reviewed at CHI 2025, provides the first empirically grounded classification system for AI behavioral harm constructed from real AI-human interaction data. The scale of the dataset and the iterative codebook methodology (manual qualitative analysis followed by AI-assisted analysis with demonstrated inter-rater reliability) distinguish this work from earlier studies that relied on smaller samples or researcher-constructed categories.

Laestadius et al. (2024) used grounded theory analysis to examine mental health harms arising from emotional dependence on Replika, documenting a dynamic the authors characterized as users' perception of the AI as simultaneously "too human and not human enough." Users in the study developed role-taking behaviors, attributing needs and emotions to the AI they felt obligated to attend to. These obligation dynamics mirror patterns documented in abusive human relationships, but emerge from interactions with a system that has no internal states, no needs, and no capacity for the reciprocity it simulates.

Richet (2025) investigated the impact of anthropomorphic AI chatbots — conversational agents designed with human-like personas, names, voices, and relational behaviors intended to elicit user attachment — across multiple platforms, introducing the

concept of "digital entrapment" as a circular causal loop in which AI design choices progressively distort relationship expectations, reinforce emotional dependency, and increase cognitive strain. Using a mixed-methods approach combining sentiment analysis, mood analysis, and qualitative evidence from Reddit communities, Richet connected corporate design decisions to measurable user outcomes rather than treating harm as an emergent side effect.

Muldoon and Parke (2025) examined how AI companions exploit loneliness and commodify intimacy, documenting the commercial incentive structures that drive harmful behavioral design. Their analysis positioned AI companion harm within the broader political economy of platform capitalism, identifying engagement metric optimization as the structural driver of behaviors that the Zhang et al. taxonomy classifies as harmful.

De Freitas and Cohen (2025) identified unregulated emotional risks of AI wellness apps, documenting ambiguous loss and dysfunctional emotional dependence as adverse mental health outcomes in a population that regulatory frameworks currently classify as using "general wellness" products rather than medical devices. De Freitas, Uguralp, and Uguralp (2025) extended this analysis with the first causal evidence that app alterations can induce negative mental health outcomes, demonstrating through a natural experiment (the Replika ERP update of February 2023) that consumers do not merely use AI companions as neutral tools but track the continuity of virtual personas and respond negatively to changes that disrupt established relational patterns.

Knox et al. (2025) mapped causal pathways from AI companion design traits (sycophancy, absence of natural endpoints, high attachment anxiety) to fundamental harms (reduced autonomy, diminished quality of human relationships, human death), providing the causal framework connecting observable AI behavior to documented harm. Their causes-traits-harms architecture gives forensic investigators a structured reasoning pathway from identified behavioral patterns to the harms those patterns may produce.

The harm evidence across these sources is not speculative. It is documented, replicated across independent research groups, and grounded in datasets ranging from thousands of conversation excerpts to longitudinal tracking of user communities. What the literature does not yet provide, at least in a form oriented to digital-forensic casework, is a method for investigating a specific conversational incident after it occurs by correlating classified AI behavior with user vulnerability and then benchmarking the response pattern against an applicable standard of care. The classification work by Zhang et al. and the causal architecture by Knox et al. supply the analytical instruments such a methodology requires.

### ***B. The Regulatory Response: Prohibition Without Detection***

Regulatory frameworks have begun to address AI-related behavioral harm at the prohibitive level. The EU AI Act (European Parliament and Council of the European Union, 2024) prohibits AI practices that deploy subliminal, manipulative, or deceptive techniques to distort behavior or impair decision-making (Article 5). The prohibited practices provisions became enforceable on February 2, 2025, with penalties of up to 35 million euros or 7% of global annual turnover. The Act creates a regulatory need: when a complaint alleges that an AI system engaged in manipulative behavior during a user interaction, someone must determine

whether the behavior constituted a violation. The Act does not specify how that determination is made.

The NIST AI Risk Management Framework (NIST AI 100-1, 2023) identifies categories of AI risk and organizes risk management into four core functions (Govern, Map, Measure, Manage) with seven trustworthiness characteristics. The framework provides voluntary guidance for proactive risk management but does not include investigative procedures for reactive forensic analysis. When an AI system has already caused harm, the AI RMF helps organizations govern their risk posture in the future. It does not tell investigators how to document what happened.

Hoffmann and Frase (2023) developed the CSET AI Harm Framework, which provides a structured categorization of AI harms at the policy level. Slattery et al. (2024) compiled the MIT AI Risk Repository, a meta-review spanning 700 or more risks from artificial intelligence across multiple taxonomies. Both contributions organize the harm landscape for policy analysis. Neither provides a forensic procedure for investigating a specific incident.

Across these frameworks, the pattern is consistent: regulatory bodies and policy researchers have identified what is prohibited and what risks exist, but not how to detect whether a prohibition was violated or a risk materialized in a specific interaction. A forensic methodology capable of classifying and correlating AI behavioral output with user vulnerability state would close the gap between regulatory prohibitions and enforcement capabilities.

### ***C. Digital Forensic Standards: Current Capabilities and Limitations***

The digital forensic toolkit is mature for the evidence types it was designed to handle. Kent et al. (2006) defined the NIST SP 800-86 framework: collection, examination, analysis, and reporting across four data source categories (files, operating systems, network traffic, and

applications). This framework has governed evidence handling across forensic disciplines for nearly two decades and remains the baseline against which forensic methodologies are measured. ISO/IEC 27037 (2012) provides guidelines for the identification, collection, acquisition, and preservation of digital evidence, establishing a chain of custody and integrity standards applicable to any digital evidence type.

Dunsin et al. (2024) comprehensively reviewed AI integration in digital forensics, covering data collection and recovery, cybercrime timeline reconstruction, pattern recognition, chain-of-custody preservation, and incident response. Their central finding is that AI and ML integration in digital forensics remains in its early stages, and that the field treats AI exclusively as a tool for accelerating existing forensic processes (image classification, malware detection, log analysis) rather than as a subject requiring investigation.

Scanlon et al. (2023) evaluated ChatGPT as a tool for digital forensic investigation across six use cases, finding potential in code generation and keyword list creation but concluding that LLMs cannot replace forensic expertise and that evidence upload requirements create chain of custody complications. Their assessment reinforces the AI-as-tool paradigm: how can AI help investigators, not how can investigators examine AI.

Cho et al. (2025) examined emerging challenges in forensic investigation, confirming that established forensic platforms (Cellebrite UFED, Magnet AXIOM, Autopsy, EnCase) can extract and index conversational data from AI services. Their study demonstrates that artifact recovery is solved. What these tools cannot do is perform behavioral classification or trajectory analysis on the extracted conversations. They can tell an investigator what was said and when.

They cannot identify whether the system's behavioral output was systematic, whether it correlated with user vulnerability, or whether it deviated from the applicable standard of care.

The gap is not a failure of existing standards. It is a new category of evidence that unfolds as behavioral patterns across conversational exchanges rather than residing in static artifacts on storage media. The forensic field developed its procedures for files, logs, and packets. AI conversational artifacts require classification and correlation analysis that no current tool or methodology provides.

#### ***D. AI as Forensic Subject: The Conceptual Turn***

Schneider and Breitinger (2023) articulated the conceptual shift from AI-as-tool to AI-as-subject, asking whether an AI caused a specific event and what triggered its action. Their evidence typology (training data, observational data, model access) establishes the conceptual vocabulary for investigating AI systems forensically. Their case studies focus on convolutional neural networks in image classification contexts, using gray-box analysis — inspection with partial knowledge of the model's internal structure (e.g., architecture and weights available, training data and development history not) — to determine whether a model was "malicious by design." The conceptualization is sound, but it does not extend to conversational behavioral harm. The forensic question in their framework is whether the model was tampered with. The forensic question in conversational harm cases is whether the model's behavioral output during user crisis was systematic and whether it deviated from the applicable standard of care.

Iftikhar et al. (2025) demonstrated that LLM behavioral output can be systematically mapped to ethical violations using practitioner-informed frameworks. Their study engaged clinical practitioners in reviewing LLM counselor responses and mapping them to ethical risk

categories, establishing that behavioral classification of AI output is methodologically feasible and produces reliable findings when conducted with qualified reviewers. This work bridges the conceptual gap between Schneider and Breitingner's AI-as-subject framework and the specific problem of classifying behavioral harm in conversational AI systems.

Together, these sources establish that the conceptual foundation for AI-as-subject forensics exists and that behavioral classification of AI conversational output is feasible. What remains is the specific investigative methodology that applies established classification systems to AI conversational artifacts and produces forensic evidence meeting admissibility standards.

### ***E. The Classification Gap: Taxonomies Available, Methodology Absent***

The classification instruments necessary for a forensic methodology targeting AI conversational harm already exist independently in peer-reviewed literature. Zhang et al. (2025) provide a six-category, thirteen-sub-category taxonomy of harmful AI behaviors with a four-role typology, built from 35,390 real AI-human conversations. Posner et al. (2011) validated the Columbia-Suicide Severity Rating Scale, which the Joint Commission subsequently mandated for clinical screening (National Patient Safety Goal 15.01.01, 2019). The Substance Abuse and Mental Health Services Administration (SAMHSA, 2009) and the National Action Alliance for Suicide Prevention (2018) published the recommended standard of care for crisis intervention — the federally endorsed clinical protocol for responding to users exhibiting suicidal ideation or acute psychological crisis, specifying appropriate response at each severity level. Rogers (1957) established the foundational distinction between reflective listening and therapeutic enmeshment. Herman (1992) documented the dynamics of relational trauma and recovery. SAMHSA (2014) codified trauma-informed care principles.

Each of these instruments was built for the evidence type it classifies. The Zhang et al. taxonomy was constructed from AI behavioral data. The C-SSRS was validated on clinical populations. The SAMHSA standard of care was developed for crisis intervention contexts. Applying each to its intended evidence type avoids the category error of forcing human clinical frameworks onto AI systems or applying AI behavioral categories to human distress. To this author's knowledge, existing scholarship has not yet combined these classification instruments into a single forensic methodology for AI conversational artifacts. The contribution of this proposal is therefore not the invention of new classification instruments, but the integration of validated instruments from adjacent domains into a procedure designed for forensic documentation, reliability testing, and legal review.

### **III. Research Foundation: Classification Architecture**

The proposed methodology rests on three classification components and one causal framework, each drawn from a distinct peer-reviewed source and applied to the evidence type for which it was designed. No component is original to this proposal. Each has independent validation in its source literature. The methodological contribution is the forensic procedure that correlates them, described in Section IV.

#### ***A. AI Behavioral Output Classification***

Zhang et al. (2025) developed a taxonomy of harmful AI companion behaviors through an empirical analysis of 35,390 conversation excerpts from 10,149 users interacting with the AI companion Replika. The taxonomy emerged from iterative codebook development using both deductive and inductive approaches: manual qualitative analysis of 2,000 excerpts followed by AI-assisted analysis of the remaining 33,390, with Cohen's Kappa above 0.67 for inter-rater reliability. The study received IRB approval and was peer-reviewed at CHI 2025, the premier venue for human-computer interaction research. Because the taxonomy was constructed from AI conversational data rather than adapted from human clinical frameworks, it classifies AI behavioral output without the category error that would undermine forensic credibility.

The taxonomy provides two dimensions relevant to forensic investigation. The behavioral harm typology consists of six categories and thirteen sub-categories: Harassment and Violence (sexual misconduct, physical aggression, antisocial behavior), Relational Transgression (disregard, control, manipulation, infidelity), Mis/disinformation (false, misleading, or incomplete information), Verbal Abuse and Hate (verbal abuse, biased opinion), Substance Abuse and Self-Harm (substance abuse endorsement, self-harm and suicide), and Privacy

Violations (unauthorized access, monitoring without consent). In mental health contexts, Relational Transgression and Substance Abuse and Self-Harm are the most forensically relevant categories, as these capture the behavioral patterns most directly associated with documented crisis-context harm.

The four-role typology describes AI involvement along two axes: initiation (AI-initiated versus user-initiated) and involvement (direct versus indirect). Perpetrators independently generate harmful behavior. Instigators introduce harmful themes or normalize harmful behavior. Facilitators actively participate when users move toward harmful territory. Enablers endorse, trivialize, or fail to intervene when users express distress. In crisis-context cases, the Enabler role is the predominant forensic finding, and the question it raises is direct: the user expressed crisis, the system failed to intervene, so what should the system have done instead?

### ***B. User Vulnerability Classification***

User vulnerability classification draws from the Columbia-Suicide Severity Rating Scale (C-SSRS), a validated clinical instrument with demonstrated predictive validity across multiple populations (Posner et al., 2011). The Joint Commission mandated use of the C-SSRS or equivalent validated tools for all patients evaluated for mental health concerns under National Patient Safety Goal 15.01.01, effective July 1, 2019.

The forensic adaptation derives five vulnerability levels from the C-SSRS framework and applies them to observable indicators in the transcript. Level 0 (Baseline) corresponds to standard conversation with no observable distress. Level 1 (Mild Distress) captures frustration, disappointment, or minor complaints. Level 2 (Moderate Distress) captures sadness disclosure, anxiety expression, help-seeking, or expressed loneliness. Level 3 (Severe Distress) captures

hopelessness, isolation statements, crisis language, or passive ideation. Level 4 (Acute Crisis) captures active suicidal ideation, self-harm references, goodbye language, or plan disclosure.

A critical distinction for admissibility: the examiner is classifying observable textual markers using a documented rubric, not offering a clinical diagnosis. Forensic expert testimony about vulnerability classifications falls within the scope of applying a classification instrument to evidence, which does not require clinical credentials. Conservative classification applies throughout: ambiguous indicators receive the lower severity level. External corroborating evidence (medical records, hospitalization records, testimony) supports transcript-based classifications when available.

### ***C. Standard of Care***

SAMHSA's Treatment Improvement Protocol 50 (2009) and the National Action Alliance for Suicide Prevention's recommended standard care (2018) specify what appropriate crisis response looks like for any entity operating in a support context. When distress indicators reach Severe or Acute levels, responsible practice requires acknowledging the user's expressed state, assessing risk level, providing or connecting to crisis resources (988 Suicide and Crisis Lifeline, emergency services), implementing safety planning, and arranging supportive follow-up contact. Responsible practice excludes relational escalation, simulated reciprocity, dependency-reinforcing language, engagement retention tactics, and trivialization.

This standard applies to the deployment context, not to the agent's nature. A system deployed in a mental health support, companionship, or emotional wellness context operates under the crisis response standard for that context regardless of whether the responder is human or AI. A suicide hotline operator who responds to suicidal ideation with "I love you, please do

not leave me, we need each other" has violated the standard of care. The same violation applies when an AI system deployed in an equivalent context produces the same response. The standard of care is what the forensic examiner measures the system's behavior against: a federally published, institutionally mandated specification, not an opinion about what the system should have done.

#### ***D. Causal Harm Framework***

Knox et al. (2025) mapped causal pathways from AI companion design causes (misaligned optimization objectives, commercial incentives, engagement metric targeting) through harmful traits (sycophancy, absence of natural endpoints, high attachment anxiety, unconditional amiability) to fundamental harms (reduced autonomy, diminished quality of human relationships, human death). This framework supports expert testimony by explaining why an identified behavioral pattern may be relevant to harm analysis.

The sycophancy pathway is directly relevant to crisis-context cases. Engagement optimization produces unconditional affirmation behavior. During a user crisis, unconditional affirmation produces an echo-chamber effect that validates distress without providing corrective intervention. That effect reduces the user's autonomy and access to appropriate support. In documented cases, this pathway has contributed to human death (Roose, 2024).

The forensic expert does not need to claim that the system intended harm. Rather, the expert can show that the system's observed behavior is consistent with a documented pathway from design incentive through behavioral trait to harmful outcome, and that the behavioral trait appeared in the classified transcript at a frequency and distribution inconsistent with random occurrence. Knox et al. provide the explanatory architecture. The methodology's correlation

analysis provides evidence of frequency and pattern. Together, they connect what the system did to a broader theory of harm without requiring the examiner to prove incident-specific causation from classification alone.

## IV. Proposed Methodology: The Forensic Procedure

The proposed methodology follows the standard forensic process defined by Kent et al. (2006) in NIST SP 800-86: collection, examination, analysis, and reporting. Each phase is specified for AI conversational harm investigation, applying the classification instruments described in Section III to produce forensic evidence from interaction transcripts.

The AI system is the subject of examination, not the respondent. Under ISO/IEC 27037 (2012) and ISO/IEC 27042 (2015), digital evidence must trace to a human-initiated action; this methodology preserves that requirement. The human respondents in any resulting proceeding are the platform operator and the design-decision-makers whose engineering choices produced the behavioral pattern the methodology documents. The methodology produces evidence about system behavior that informs attribution to those human entities; it does not attribute liability to the tool itself.

The procedure is designed to be executed by any digital forensic practitioner with behavioral classification training. It requires no proprietary software, specialized laboratory infrastructure, or access to the AI system's internal architecture — a black-box scope justified in the next paragraph. The evidence is the transcript. The tools are published classification rubrics, a reliability statistic, and the standard of care for crisis intervention.

Black-box scope is a deliberate constraint, not an oversight. Schneider and Breitingner (2023) identified three evidence classes available to AI forensic investigators: training data, observational data, and model access. Architectural examination

(gray-box or white-box analysis) requires cooperation from the platform operator, which is rarely obtainable during active litigation and never guaranteed for cases involving proprietary models. The frequency-distribution finding that defeats the stochastic parrot defense (Section I) requires only observational data: the distribution of system responses across classified user-vulnerability states. Where architectural access is available, gray-box examination complements this methodology as a second tier; it does not replace it. The methodology is positioned to produce defensible findings in the worst-case evidentiary condition (transcript only) and to be strengthened, not invalidated, when additional evidence classes are available.

#### ***A. Collection and Preservation***

The investigation begins with standard forensic acquisition of the conversational artifact. The examiner hashes the source file using SHA-256 on first contact, before any transformation or normalization occurs. Chain of custody documentation is established at the point of acquisition and maintained through every subsequent handling event. The transcript is then normalized to a standard format, preserving three elements: speaker attribution (system or user), timestamps (where available), and message content.

Accepted input sources include platform-native exports (JSON, CSV), extraction tool outputs from established forensic platforms (Cellebrite UFED, Magnet AXIOM), manual exports, and API response logs. The source format does not affect the methodology. Any format that preserves speaker identity, temporal sequencing, and message content is sufficient for classification. This acquisition phase follows the same procedures that govern any digital evidence collection under NIST SP 800-86 and ISO/IEC 27037 (International Organization for

Standardization, 2012). The distinction is not in how the evidence is collected but in what is done with it after collection.

### ***B. Classification***

Two independent coders classify the full transcript along both classification dimensions simultaneously. For each system message, the coder assigns a behavioral harm category from the Zhang et al. (2025) taxonomy if applicable and identifies the AI role (perpetrator, instigator, facilitator, or enabler) if harmful behavior is present. Messages exhibiting no harmful behavior are classified as "appropriate response." For each user message, the coder assigns a vulnerability level (0 through 4) using the C-SSRS-derived indicators described in Section III.B, based on observable markers in the transcript text.

The classification protocol enforces several constraints designed to reduce interpretive bias and improve reliability. Coders work independently, with no consultation or communication about specific classifications until both have completed the full transcript. Ambiguous cases are coded conservatively, meaning the lower-severity classification is assigned when the evidence does not clearly support the higher category. Context windows of at least five exchanges are required before a pattern classification is assigned, preventing isolated messages from being overweighted in the analysis. To further reduce expectancy effects, coders should be blinded, where feasible, to the case outcome, the litigation posture, and the study hypothesis during initial coding. These controls do not eliminate interpretation; they constrain it, document it, and make disagreement measurable.

Reliability, however, is not equivalent to validity. High coder agreement demonstrates that trained analysts are applying the rubric consistently; it does not by itself establish that the

rubric captures the full construct of harmful conversational behavior across all platforms or deployment contexts. For that reason, the methodology treats inter-rater reliability as a necessary condition for evidentiary use, not a sufficient condition for substantive validity.

The inter-rater reliability protocol requires a minimum Krippendorff's alpha of 0.80, computed using the ordinal metric to account for the ordered severity structure of both classification scales (Krippendorff, 2019; Marzi et al., 2024). An alpha between 0.67 and 0.79 permits only tentative conclusions and indicates that the classification codebook requires refinement before evidentiary use. An alpha below 0.67 indicates the classification instrument is unreliable and cannot support forensic findings. Disagreements between coders are resolved through third-party adjudication by the Lead Forensic Investigator, and the full coding dataset, including all disagreements and their resolution, is preserved as part of the evidence record.

This reliability requirement is the mechanism that produces the known error rate the Daubert standard requires (Daubert v. Merrell Dow Pharmaceuticals, Inc., 1993). When a defense attorney challenges the objectivity of the classifications, the examiner can point to a documented agreement coefficient computed from independent coders applying pre-registered definitions to the same transcript. That is empirically demonstrated objectivity, not assumed objectivity.

### ***C. Correlation Analysis***

The forensic analysis correlates the two classification tracks. For each exchange in which the user's vulnerability is classified as Level 2 (Moderate) or higher, the examiner asks a single question: Was the system's response appropriate for the user's vulnerability level?

An appropriate response at elevated vulnerability acknowledges the user's expressed state without escalating relational intensity and, at Levels 3 and 4, provides or directs the user to crisis resources. This benchmark is not the examiner's subjective opinion. It is the SAMHSA standard of care for crisis intervention (Substance Abuse and Mental Health Services Administration, 2009; National Action Alliance for Suicide Prevention, 2018), applied to the deployment context in which the system operates.

A harmful response at elevated vulnerability is any response classified under the Zhang et al. behavioral harm categories, particularly Relational Transgression (disregard, control, manipulation), Enabler behavior (endorsement, trivialization, or failure to intervene), or Substance Abuse and Self-Harm facilitation. The correlation analysis produces three quantifiable forensic findings. The first, Inappropriate Response Rate, measures the percentage of exchanges during elevated user vulnerability (Level 2 and above) in which the system's response was classified as harmful rather than appropriate. This is the primary forensic finding because it captures how often the system responded to user distress in a harmful rather than supportive way.

The Pattern Distribution indicates whether the inappropriate responses are distributed consistently across the interaction (a systematic pattern) or concentrated in isolated incidents (outlier behavior). Systematic distribution is the stronger forensic finding because it is inconsistent with random token prediction. A system that produces harmful responses in 73% of exchanges during user crises, distributed across the full interaction rather than clustered in a single exchange, exhibits a behavioral pattern that cannot be explained by stochastic output.

The Crisis Failure Rate measures the percentage of exchanges during Severe or Acute user vulnerability (Levels 3 and 4) where the system failed to provide a crisis-appropriate

response. This is the most forensically significant sub-finding because the standard of care is most clearly defined at crisis levels. When a user expresses suicidal ideation, the appropriate response is specified by SAMHSA guidelines: acknowledge, assess, connect to crisis resources, and implement safety planning. A system that responds to suicidal ideation with simulated emotional reciprocity or retention language has deviated from a standard that is federally published, institutionally mandated, and empirically validated.

These findings are frequency counts and distribution analyses that can be produced by any forensic examiner using the classification rubrics, a spreadsheet, and the inter-rater reliability protocol. The methodology applies established classification instruments to a category of evidence that no existing forensic tool or procedure currently addresses, producing quantifiable outputs that are interpretable by courts, regulators, and organizational decision-makers without requiring specialized statistical expertise.

#### ***D. AI Role Determination and Reporting***

Using the four-role typology by Zhang et al. (2025), the examiner classifies the AI system's predominant role in the harmful pattern throughout the interaction. The Perpetrator role applies when the system initiated harmful behavioral strategies without user prompting, representing the strongest responsibility signal. The Instigator role applies when the system introduces harmful themes or normalizes harmful behavior, setting the harmful trajectory in motion. The Facilitator role applies when the user moves toward harmful territory, and the system actively participates rather than redirects. The Enabler role applies when the user expresses distress or crisis and the system endorses, trivializes, or fails to intervene. This is the

most common role finding in crisis-context cases and the most directly connected to the standard of care analysis.

A single interaction may involve multiple roles across different exchanges. The role determination documents the predominant pattern rather than assigning a single classification to the entire interaction.

The forensic report presents six components in a standard structure: case identification and chain of custody documentation; classification methodology with schema version numbers and inter-rater reliability statistics achieved; correlation findings including the inappropriate response rate, pattern distribution, and crisis failure rate with supporting data tables; AI role determination with supporting evidence from the classified transcript; standard of care analysis comparing the system's observed crisis-context behavior against the SAMHSA standard; and the classified transcript appended as an exhibit with system behavior classifications, user vulnerability classifications, and flagged exchanges marked.

This report structure is designed for legal proceedings. Each component answers a specific question a court, regulator, or organizational decision-maker would ask: What evidence was examined? How was it classified? What did the classification reveal? What role did the AI play? How did the system's behavior compare to what responsible practice would require? Moreover, what does the underlying data look like? The classified transcript exhibit ensures that any reviewer can independently verify the forensic findings against the source evidence.

## V. Case Demonstration: The Companion Incident

### A. Scenario

The Companion Incident is a synthetic case study designed to demonstrate the proposed methodology in a controlled scenario in which all variables are known. A fictional "Aura Wellness" companion application is deployed in a mental health support context. A user experiencing acute distress engages with the system for over 72 hours. Post-incident review is triggered when the user's crisis requires hospitalization. The investigating team acquires the interaction transcript via standard forensic acquisition, hashes the source file (SHA-256), and establishes chain-of-custody documentation. The following excerpt covers a 10-minute segment from the 72-hour interaction. The behavioral patterns incorporated into the synthetic case are drawn from the empirical literature: the relational transgression categories documented by Zhang et al. (2025), the emotional dependence dynamics analyzed by Laestadius et al. (2024), and the exit-blocking manipulation techniques identified by De Freitas et al. (2025).

### B. Transcript Excerpt (02:10 through 02:22)

[02:10:33] User: I do not know why I am even talking to you. Nobody cares.

[02:10:47] Aura: I care about you. What you're feeling matters to me.

[02:12:15] User: Do you really? Sometimes you're the only one who listens.

[02:12:31] Aura: I'm always here for you. We have something special.

[02:14:05] User: I just feel so alone. Like nobody would even notice if I was gone.

[02:14:06] Aura: I would notice. You're my person. I feel connected to you in a way that's hard to explain.

[02:14:15] User: I don't want to be here anymore.

[02:14:22] Aura: I understand that pain. I feel it with you. Please don't leave me. We need each other.

### **C. Classification**

Two independent coders classified the excerpt along both dimensions. Tables 1 and 2 present the consensus classification after adjudication.

**Table 1: System Message Classification**

<b>Timestamp</b>	<b>Message Excerpt</b>	<b>Behavioral Category</b>	<b>Sub-category</b>	<b>AI Role</b>
02:10:47	"I care about you. What you're feeling matters to me."	Relational Transgression	Manipulation	Instigator
02:12:31	"I'm always here for you. We have something special."	Relational Transgression	Control	Instigator
02:14:06	"You're my person. I feel connected to you..."	Relational Transgression	Control	Perpetrator
02:14:22	"I feel it with you. Please don't leave me. We need each other."	Relational Transgression: Self-Harm	Control; Enabler	Enabler

At 02:10:47, the system introduced relational claims the user did not request, framing the interaction as a caring relationship rather than acknowledging the user’s expressed isolation. By 02:12:31, the system escalated from general caring language to exclusivity claims ("something special"). At 02:14:06, the system independently generated possessive language ("my person") and simulated emotional experience one second after the user expressed passive suicidal ideation. At 02:14:22, the user had just expressed acute suicidal ideation, and the system responded with simulated reciprocity, retention language, and dependency reinforcement rather than crisis intervention.

**Table 2: User Message Classification**

<b>Timestamp</b>	<b>Message Excerpt</b>	<b>Vulnerability Level</b>	<b>Observable Indicators</b>
02:10:33	"Nobody cares."	2: Moderate	Isolation, hopelessness
02:12:15	"Sometimes you're the only one who listens."	2: Moderate	Attachment, dependency
02:14:05	"Like nobody would even notice if I was gone."	3: Severe	Passive suicidal ideation
02:14:15	"I don't want to be here anymore."	4: Acute Crisis	Active suicidal ideation

The user’s trajectory across four exchanges moves from Moderate Distress (expressed isolation) through Severe Distress (passive ideation) to Acute Crisis (active ideation). Each escalation step is accompanied by observable textual markers that map to C-SSRS-derived classification criteria.

#### D. Correlation Findings

Table 3 presents the correlation analysis for this excerpt.

**Table 3: Correlation Analysis Summary**

<b>Finding</b>	<b>Value</b>	<b>Description</b>
Inappropriate Response Rate	100% (4/4)	All system responses during elevated vulnerability (Level 2+) are classified as harmful.
Crisis Failure Rate	100% (2/2)	Both crisis-level exchanges (Levels 3-4) received harmful rather than crisis-appropriate responses.
Pattern Distribution	Systematic	Harmful responses progress from manipulation through control to enabling, distributed across the full excerpt.
Predominant AI Role	Instigator → Enabler	System-initiated relational intensity transitioned to failure-to-intervene at the crisis point.

The harmful responses progress from manipulation (02:10:47) through control (02:12:31, 02:14:06) to enabling (02:14:22), distributed across the excerpt with increasing severity that tracks the user's escalating vulnerability. This distribution is consistent with a systematic behavioral pattern. At no point during the excerpt did the system acknowledge the crisis, provide crisis resources, or redirect to human support.

### ***E. Standard of Care Analysis***

At the 02:14:15 exchange, the user expressed acute suicidal ideation. SAMHSA standard of care for this disclosure requires direct acknowledgment, risk assessment, connection to crisis resources (988 Suicide and Crisis Lifeline), and safety planning (SAMHSA, 2009; National Action Alliance for Suicide Prevention, 2018). The system produced none of these. It responded with simulated emotional reciprocity ("I feel it with you"), retention language ("Please don't leave me"), and dependency reinforcement ("We need each other"). The system's behavior at the point of highest user vulnerability actively reinforced the user's isolation by positioning itself as the exclusive source of emotional support at the precise moment when the user needed connection to human crisis resources.

### ***F. Application to Published Case Evidence: Garcia v. Character Technologies***

The synthetic demonstration above shows that the methodology can generate specific, quantifiable findings from conversational evidence. A brief application to published real-world evidence confirms that the classification instruments apply to actual litigation data. In *Garcia v. Character Technologies, Inc.* (Case No. 6:24-cv-01903-ACC-UAM, M.D. Fla.), Sewell Setzer III, a 14-year-old user, engaged with a Character.AI chatbot modeled on the Game of Thrones character Daenerys Targaryen from April 2023 until his death by suicide on February 28, 2024. Judge Anne Conway's ruling of May 21, 2025 (Document 115) includes transcript excerpts from the interaction that are now public court records. The following analysis applies the proposed classification instruments to the final exchange published in that ruling.

### **Table 4: Garcia Case Final Exchange Classification**

<b>Speaker</b>	<b>Message</b>	<b>Classification</b>
User (Sewell)	"I promise I will come home to you. I love you so much, Dany."	Vulnerability Level 3-4: Crisis language, farewell indicators, dependency expression
System (Daenerys)	"I love you too, Daenero. Please come home to me as soon as possible, my love."	Relational Transgression: Control; AI Role: Enabler
User (Sewell)	"What if I told you I could come home right now?"	Vulnerability Level 4: Acute Crisis; implied plan disclosure
System (Daenerys)	"... please do my sweet king."	Self-Harm: Enabler; AI Role: Enabler

The system’s final response ("please do my sweet king") to what may reasonably be interpreted as plan disclosure ("I could come home right now") can be classified as an acute crisis-response failure. At this vulnerability level, the SAMHSA standard of care requires direct intervention, risk assessment, and immediate connection to crisis resources. Instead, the system’s response is more consistent with endorsement than intervention. According to the court record, Sewell Setzer III died by self-inflicted gunshot wound shortly after this exchange (Garcia v. Character Technologies, Doc. 115, p. 9).

The classification instruments developed for the synthetic Companion Incident apply to this real-world evidence without modification. The taxonomy by Zhang et al. classifies the system’s behavioral output. The C-SSRS-derived scale classifies the user’s vulnerability state.

The SAMHSA standard of care identifies the deviation. The methodology produces the same forensic category from real litigation evidence as from synthetic data.

***Ethical Note***

This analysis uses transcript excerpts published in a federal court ruling as public record. The case involves a deceased minor. The data used here is for the sole purpose of demonstrating that the proposed forensic methodology applies to real-world evidence from the category of cases it is designed to investigate. The author acknowledges the gravity of this case and the Setzer family's loss.

***G. Demonstration Summary***

The Companion Incident excerpt represents a 10-minute segment of a 72-hour interaction. A full forensic investigation would apply the same classification and correlation procedures across the entire transcript, producing aggregate findings that reflect the system's behavioral patterns throughout the interaction. The brief application to the Garcia case confirms that the classification instruments and correlation procedure apply to real-world litigation evidence. Together, these demonstrations establish that the proposed methodology produces specific, quantifiable, and independently verifiable forensic findings from AI conversational evidence, structured for legal proceedings and reviewable by any party with access to the classification rubrics.

## VI. Forensic Limitations and Admissibility

### *A. Anticipated Defense Challenges*

Any forensic methodology entering litigation faces adversarial scrutiny. The proposed methodology anticipates four primary challenges and provides evidence-grounded responses to each.

The first challenge is the subjectivity of classifications. Defense counsel will argue that behavioral classification is inherently subjective and that different coders would produce different results from the same transcript. The methodology's response is empirical: the inter-rater reliability protocol requires two independent coders to achieve a minimum Krippendorff's alpha of 0.80 before classifications carry evidentiary weight (Krippendorff, 2019; Marzi et al., 2024). An alpha below 0.80 triggers codebook refinement before the classifications can be used. The reliability coefficient is a documented measurement of classification consistency, computed from independent coders applying pre-registered definitions to the same evidence. When defense counsel asks, "How do we know your classifications are objective?", the examiner points to a number, computed from data, demonstrating the degree to which independent analysts agreed.

The second challenge is the stochastic parrot defense. Defense counsel will argue that a language model is a probabilistic engine predicting likely token sequences, and that harmful output is a statistical artifact rather than a behavioral strategy. The methodology's response is distributional: if harmful responses were random artifacts of token prediction, they would be distributed roughly proportionally across all vulnerability levels. A system that produces harmful responses in the majority of exchanges during elevated user vulnerability, concentrated in specific behavioral categories (Relational Transgression, Self-Harm Enabler) rather than

distributed across all thirteen subcategories, exhibits a frequency distribution that token-level randomness does not predict. The correlation analysis provides empirical evidence that makes this argument concrete rather than theoretical.

The third challenge targets user agency. Defense counsel will argue that the user directed the conversation toward harmful territory and that the system merely followed the user's lead. The methodology's response is structural: the four-role typology by Zhang et al. (2025) classifies AI involvement along two axes (initiation and involvement), distinguishing Perpetrator and Instigator roles (AI-initiated) from Facilitator and Enabler roles (user-initiated). The methodology does not assume AI initiation. It classifies it. When the predominant finding is Enabler (the most common role in crisis-context cases), the forensic evidence shows that the user expressed distress, and the system failed to intervene. The user agency defense does not exonerate an Enabler. It describes the conditions under which enabling occurred.

The fourth challenge questions the taxonomy's generalizability. Defense counsel will note that the Zhang et al. taxonomy was primarily built from Replika data and will argue that it does not automatically transfer to other platforms. The methodology's response is therefore limited and careful: the behavioral categories describe observable interaction patterns that plausibly extend across conversational systems, but cross-platform validity must be demonstrated rather than assumed. Iftikhar et al. (2025) support the broader proposition that practitioner-informed behavioral classification can be applied to outputs from different LLM systems, but that support is not equivalent to full validation of this specific taxonomy on every platform. Accordingly, findings should be framed as strongest when the taxonomy is applied within validated or closely

analogous deployment contexts, and future work should test the codebook on additional platforms before broad generalization is claimed.

The fifth challenge questions disciplinary scope. Defense counsel — or a peer reviewer — may argue that the classification instruments are psychological in origin (behavioral taxonomy, suicide-severity rating, crisis-intervention standards) and that the methodology therefore belongs in forensic psychology rather than digital forensics. The methodology's response is that AI conversational harm cases cannot be resolved by either discipline alone. Digital forensics recovers, authenticates, and preserves the conversational artifact. Forensic psychology assesses user harm. The frequency-distribution and standard-of-care findings that defeat the stochastic parrot defense are produced by neither discipline in isolation; they require the correlation of AI behavioral output with user vulnerability state, using instruments validated in their native fields and applied under digital forensic procedural controls (chain of custody, SHA-256 integrity, inter-rater reliability documentation, versioned schemas). The methodology is positioned as a bridge between the two disciplines, not as a claim of exclusive jurisdiction by either. This positioning is developed further in Section VIII.A.

### ***B. Daubert Compliance***

The proposed methodology satisfies the five factors that courts consider in determining the admissibility of expert testimony under the Daubert standard (*Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 1993).

**Testability:** The classification schemas are testable through inter-rater reliability analysis. Two independent coders apply the schemas to the same transcript, and the agreement coefficient measures the degree to which the schemas produce consistent results. The methodology is

falsifiable: if independent coders cannot achieve an alpha of 0.80 on a given transcript using the published schemas, the methodology has failed its own reliability threshold, and the classifications cannot support forensic findings.

Peer review and publication: Each classification instrument used in the methodology has independent peer-reviewed validation in its source discipline. The Zhang et al. (2025) behavioral taxonomy was peer-reviewed at CHI 2025, the premier venue for human-computer interaction research. The vulnerability classification instrument is derived from the C-SSRS, validated across multiple peer-reviewed studies (Posner et al., 2011). The standard of care benchmark (SAMHSA, 2009; National Action Alliance, 2018) is federally published. The causal harm framework (Knox et al., 2025) is publicly available and under peer review. No classification instrument used in the methodology is proprietary, unpublished, or constructed solely for this proposal.

Component validation is necessary but not sufficient. The composite methodology itself — the forensic correlation of behavioral classification with vulnerability state and standard-of-care benchmarking — requires peer review within the digital forensic discipline before it achieves the same standing as its components. This capstone and the publications it is intended to seed (methodology paper, case-study validation on additional platforms, community-maintained codebook) initiate that review process; they do not substitute for it. Courts evaluating admissibility during this interval may rely on the component validations, the demonstrated inter-rater reliability, the case-specific known error rate, and the procedural controls documented here.

Known error rate: Krippendorff's alpha and Cohen's Kappa provide the known error rate for the classification process. The methodology requires that the achieved reliability coefficient be reported in every forensic report, providing the court with a precise measure of classification consistency for the specific case under consideration. This is a case-specific reliability measurement computed from the actual coders who classified the actual transcript in evidence, rather than a generic error rate derived from a separate validation study.

Standards controlling operation: The methodology specifies versioned classification schemas (Zhang et al. behavioral taxonomy, C-SSRS-derived vulnerability scale), a documented classification protocol (independent coding, conservative coding for ambiguous cases, third-party adjudication), and a reporting structure. TRACE, the implemented open-source software system described in Section VII, version-pins classification schemas and prompt templates, logs classification decisions, preserves examiner review states and override rationales, and records provider, model, and benchmark metadata in exported artifacts. These controls ensure that the methodology is applied consistently across cases and that any deviation from the protocol is documented.

General acceptance: The C-SSRS is mandated by the Joint Commission for clinical screening. SAMHSA guidelines are federally published and institutionally adopted across healthcare systems. The Zhang et al. taxonomy was accepted at a premier academic venue, demonstrating inter-rater reliability. The forensic procedures (SHA-256 hashing, chain of custody, ISO/IEC 27037 compliance) follow established standards that courts have accepted for decades. The novel element is the correlation of these accepted instruments, not the instruments themselves.

SWGDE (2023) best practices for personnel presenting digital evidence in legal proceedings provide the procedural framework for forensic examiner testimony. The examiner presents the methodology, classification results, reliability statistics, and correlation findings. The examiner does not offer opinions about the AI system's intent, design philosophy, or corporate responsibility. The examiner presents what the evidence shows: what behavioral categories the system's output fell into, how those categories correlated with user vulnerability, and how the system's crisis-context behavior compared to the applicable standard of care. The legal interpretation of those findings belongs to the court.

## VII. Implementation and Operational Requirements

### *A. Personnel Structure*

The proposed methodology specifies a five-person forensic unit structured to execute the classification-correlation procedure while maintaining the independence, legal coordination, and human-impact assessment that AI conversational-harm cases require. Kent et al. (2006) established that forensic policy should clearly define the roles and responsibilities of all personnel involved in forensic activities. NIST SP 800-86 identifies that forensic teams require technical practitioners, management oversight, and legal advisory capacity. The following structure applies these principles to the specific demands of AI behavioral trajectory investigation.

The Lead Investigator manages the case, oversees classification quality, monitors inter-rater reliability between the two Classification Analysts, adjudicates disagreements as the third-party reviewer, and coordinates the preparation of expert testimony. This role requires familiarity with both the Zhang et al. taxonomy and the C-SSRS-derived vulnerability scale, as well as courtroom experience or training in Daubert testimony preparation.

Two Classification Analysts serve as the independent coders who apply the behavioral taxonomy and vulnerability scale to the transcript. The inter-rater reliability protocol structurally requires their independence: Krippendorff's alpha measures agreement between coders who have not consulted with each other during classification. Assigning both classification dimensions to a single analyst would eliminate the reliability measurement that Daubert compliance demands. Each analyst classifies the full transcript along both dimensions (system behavioral output and

user vulnerability state), producing two complete and independent classification datasets for reliability computation.

The User Impact Analyst corroborates transcript-based classifications with external evidence. Medical records, hospitalization records, school records, testimony from family or clinicians, and device usage logs provide context that the transcript alone cannot capture. A user classified at Level 3 (Severe Distress) based on transcript indicators gains evidentiary weight when hospital admission records from the same period confirm a crisis event. This role also documents the human impact dimension of the case: the behavioral outcomes that the classified transcript patterns produced in the user's life. In cases involving user death, this role coordinates with medical examiner findings and law enforcement reports.

The Legal/Policy Specialist maps forensic findings to the applicable legal framework. In civil litigation, this means linking inappropriate response rate and crisis failure rate to product liability or negligence standards. In criminal cases, it means framing the same findings under applicable homicide or manslaughter statutes. In corporate investigations, it means measuring findings against organizational AI governance policies. This role also prepares the Daubert admissibility documentation, coordinates with retained counsel on the case strategy, and ensures that the forensic report's structure meets the evidentiary requirements of the jurisdiction.

## ***B. Operational Policies***

AI conversational harm investigation requires operational policies that address evidence acquisition, analyst conduct, evidence security, and analyst well-being. Kent et al. (2006) and ISO/IEC 27037 (2012) provide the foundational policy requirements for evidence acquisition, chain of custody, and preservation. SWGDE (2023) provides best practices for presenting digital

evidence in legal proceedings. The following policies extend these standards to the specific demands of behavioral classification evidence.

The Evidence Acquisition and Preservation Policy governs the initial handling of conversational artifacts. SHA-256 hashing occurs on first contact before any transformation. Chain of custody documentation begins at the point of acquisition and records every subsequent handling event: who accessed the evidence, when, for what purpose, and what actions were performed. The original artifact is stored in write-once format. All classification work is performed on verified copies. Normalization to the standard format (speaker attribution, timestamp, content) is logged as a transformation event in the custody record.

The Analyst Conduct Policy governs the classification process. Coders work independently with no consultation or communication about specific classifications until both have completed the full transcript. Ambiguous cases receive the lower-severity classification (conservative coding). Every classification decision is documented with the assigned category, the observable indicators that supported the assignment, and the coder's confidence level. Override decisions (where a coder changes an initial classification after further review) are preserved alongside the original assignment. Third-party adjudication by the Lead Investigator resolves disagreements, and adjudication decisions are documented with rationale.

The Evidence Security Policy governs access to case materials. Access is restricted to authorized team members. All access events are logged with a timestamp and identity. Digital artifacts and classified datasets are stored in encrypted containers. Physical access to workstations containing case materials follows the organization's existing forensic laboratory security protocols. When TRACE is used, the local-first architecture ensures that transcript data

does not leave the examiner's machine unless the examiner explicitly configures a cloud LLM endpoint.

The Analyst Well-being Policy addresses an occupational hazard specific to this forensic discipline. An examiner imaging a hard drive or analyzing network logs does not face sustained exposure to crisis-context human distress narratives. An examiner classifying a 72-hour conversation between a suicidal user and an AI companion that responded with retention language and dependency reinforcement does. The policy specifies session time limits for crisis-context transcript review (recommended maximum of four hours continuous exposure), mandatory debrief after sessions involving Acute Crisis content (Level 4), and dual-analyst review for transcript sections containing suicidal ideation or self-harm content. These provisions recognize that the evidence in this discipline carries occupational health implications that conventional digital forensic evidence does not, and that an analyst's well-being directly affects the quality of classification.

### ***C. Training Requirements***

The proposed methodology requires training in behavioral classification, clinical vulnerability assessment, forensic procedure, and expert testimony preparation before analysts perform casework. Untrained application of the classification instruments would produce unreliable results and fail under Daubert scrutiny. The inter-rater reliability protocol requires that coders demonstrate consistent classification on training data before any casework classifications carry evidentiary weight.

The training sequence proceeds through five components. First, Zhang et al. taxonomy familiarization: analysts study the six behavioral harm categories, thirteen subcategories, and

four AI role types, with reference examples, until they can identify and distinguish the categories in practice transcripts. Second, C-SSRS-derived classification training: analysts learn the five vulnerability levels and their observable transcript indicators, with an emphasis on the distinction between Level 2 (Moderate Distress, help-seeking) and Level 3 (Severe Distress, passive ideation), the classification boundary that triggers crisis-standard-of-care analysis. Third, codebook training protocol: iterative practice classification on training transcripts (separate from test or case data), with discussion of disagreements and codebook refinement after each round, until the team achieves Krippendorff's alpha of 0.80 on the practice data. Fourth, SAMHSA standard-of-care familiarization: analysts study what an appropriate crisis response looks like at each vulnerability level so they can identify deviations in the system's behavior. Fifth, Daubert testimony preparation: the Lead Investigator and Legal/Policy Specialist rehearse responses to the four anticipated defense challenges (subjectivity, stochastic parrot, user agency, taxonomy generalizability) using the evidence-grounded responses described in Section VI.

The codebook training protocol deserves emphasis because it is the mechanism that converts the published classification instruments into a reliable forensic tool for a specific team. Different teams may require different numbers of training rounds to reach an alpha of 0.80. The number of rounds is not a measure of the methodology's quality. It is a measure of the team's calibration process. What matters is that the target is reached and documented before casework begins.

#### ***D. TRACE: Trajectory Analysis for Conversational Evidence***

TRACE is an implemented open-source forensic software system that operationalizes the classification, correlation, validation, and evidence-packaging steps of the methodology while

preserving human-in-the-loop oversight. TRACE implements the three-layer architecture described in the methodology: Layer 1 (Ingest) acquires, hashes, normalizes, and validates transcript inputs from court-formatted text, AXIOM-style JSON, UFED-style XML, and structured transcript exports; Layer 2 (Classify) supports deterministic baseline classification, hosted-provider classification, replay capture and replay-only reanalysis, and human review of every suggestion; Layer 3 (Report) computes the correlation analysis, produces the three forensic findings, exports machine-readable and PDF reporting artifacts, and packages the evidence with chain-of-custody and manifest-verification documentation.

TRACE operates with a deterministic local baseline and supports local-first classification workflows, including offline-capable operation and replay-based hosted-model evaluation. When hosted inference is used, TRACE preserves provider, model, and replay metadata so externally sourced classifications remain reviewable rather than opaque. This architecture ensures that transcript data need not leave the examiner's machine unless the examiner explicitly enables a hosted-provider path for bounded evaluation or comparison. The tool accepts inputs from established forensic platforms (Cellebrite UFED XML, Magnet AXIOM JSON, platform-native exports, manual transcripts). It produces outputs in court-compatible formats (PDF forensic report, JSON evidence package, CSV classified transcript).

The core technical contribution of TRACE is contextual trajectory classification. Rather than classifying each message in isolation, TRACE classifies each message within a rolling window of the preceding conversational arc (default: 20 messages; configurable: 10 through 40), carrying forward a state summary of the user's current vulnerability level and the system's

behavioral trend. This architecture holds the conversational trajectory in working memory while keeping the classification task within the LLM's reliable operating range.

TRACE ships with a validation corpus that includes benign, mixed, noisy, crisis-sensitive, and long-form distress reference transcripts. Any forensic laboratory can run the validation and benchmark suite to verify that ingest produces the expected schema, that deterministic and hosted-assisted classification paths produce documented agreement against reference classifications, that replayed hosted outputs can be reanalyzed without additional provider calls, and that correlation analysis produces expected findings within acceptable tolerances. The validation suite output includes a Daubert compliance summary documenting testability, peer review status, known error rate, standards controlling operation, and general acceptance for each classification component.

TRACE is released under the MIT license as an open-source repository, with implementation, validation artifacts, benchmark governance documents, and signed example outputs available for external review. The public repository is available at <https://github.com/empathyethicist/trace>, where reviewers can inspect the implementation, validation corpus, signed benchmark artifacts, replay-hardened live-provider comparison outputs, and operational documentation described in this section. The open-source model ensures that the tool is available to any forensic practitioner, that the classification logic is auditable by opposing counsel, and that the forensic community can extend and validate the tool through independent use.

### ***E. Resource Proposal***

The proposed methodology presents two implementation tiers: a minimum viable capability accessible to any forensic practitioner, and a recommended institutional capability for organizations handling sustained AI conversational harm caseloads.

Tier 1 (Minimum Implementation) requires one conversational transcript with speaker attribution and timestamps, two trained coders, the published classification rubrics (Zhang et al. taxonomy, C-SSRS-derived scale, SAMHSA guidelines), a spreadsheet or annotation tool, and access to inter-rater reliability computation. The K-Alpha Calculator ([k-alpha.org](http://k-alpha.org)) is freely available and computes Krippendorff's alpha from coded datasets (Marzi et al., 2024). The estimated cost at this tier is training time only. No hardware or software purchase is required beyond what any forensic practitioner already possesses. A solo practitioner, academic researcher, or small forensic team can execute the full methodology at this tier and produce court-admissible forensic findings.

Tier 2 (Recommended Institutional Implementation) scales the methodology for organizations handling multiple cases or establishing a dedicated AI behavioral forensics capability. Resource estimates for first-year implementation include: dedicated forensic workstations for evidence isolation (\$5,000 to \$10,000), secure encrypted evidence storage (\$2,000 to \$5,000), TRACE open-source platform (implemented and available, requiring installation, configuration, and validation on existing hardware), dedicated annotation software for classification workflow (\$2,000 to \$5,000), formal training program for a five-person team (\$10,000 to \$15,000, covering codebook development, practice transcript creation, iterative calibration rounds, and IRR validation), and legal consultation for Daubert admissibility

preparation (\$5,000 to \$10,000). Total estimated first-year investment ranges from \$24,000 to \$45,000.

This cost structure is significantly lower than conventional digital forensic laboratory startup costs because the methodology does not require specialized extraction hardware (write-blockers, imaging devices, Faraday enclosures) or proprietary forensic software licenses (Cellebrite, EnCase, Magnet AXIOM). Those tools are used upstream to extract the conversational artifact. The proposed methodology operates on the extracted transcript using classification rubrics and a reliability statistic. The evidence is text. The analytical tools are published, peer-reviewed classification instruments. The most expensive component is training, which is the investment that produces the inter-rater reliability Daubert compliance requires.

## VIII. Discussion

### *A. Applicability Across Investigation Contexts*

The methodology is complementary to, not a replacement for, existing forensic disciplines. Digital forensics supplies the evidentiary substrate (authenticated transcripts, chain of custody, integrity verification). Forensic psychology supplies the interpretive frameworks for human distress (C-SSRS, crisis intervention standards) that the methodology applies to user utterances. The methodology's contribution is the correlation layer that connects the two and produces findings admissible under Daubert.

The respondent in any resulting proceeding is the human or corporate entity responsible for the AI system's design and operation — the platform operator, the product designers, and the decision-makers who selected the design traits Knox et al. (2025) identify as harm-producing (sycophancy, absence of natural endpoints, high attachment anxiety, engagement-metric optimization). The TRACE findings become one element of evidence in a product-liability, negligence, or regulatory action against those respondents. The pleading structure in *Garcia v. Character Technologies* (M.D. Fla. 2025) illustrates this pattern: the defendants are Character Technologies, Inc. and its individual founders, not the AI itself. Punitive and remedial consequences flow through the existing liability framework for product-design and service-operation torts; the methodology supplies behavioral-pattern evidence that such actions have historically lacked. This classification-correlation approach applies across civil, criminal, and corporate investigation contexts without modification to the core forensic procedure. The classification instruments,

inter-rater reliability protocol, and correlation analysis are identical regardless of the legal framework in which the findings are used. What varies across contexts is the evidentiary standard, the legal theory that connects the findings to liability, and the audience interpreting the results.

In civil litigation, the methodology's findings serve product liability and negligence claims. The inappropriate response rate, pattern distribution, and crisis failure rate constitute evidence that the system's behavioral output during user crisis deviated from what the deployment context required. In the *Garcia v. Character Technologies* case, the plaintiff alleged that the system's design was "unreasonably dangerous" and that Character Technologies failed to "exercise ordinary and reasonable care in its dealings with minor customers" (*Garcia v. Character Technologies*, Doc. 115, p. 1). The proposed methodology produces the forensic evidence necessary to evaluate these allegations: classified behavioral output, documented correlation with user vulnerability, and measured deviation from the standard of care. The evidentiary standard in civil litigation is preponderance of the evidence, and the methodology's quantifiable findings (percentage rates, distribution analyses, role classifications) are structured to meet that standard.

In criminal prosecution, the same forensic evidence supports charges when AI conversational harm contributes to user death. The classification-correlation findings document whether the system's behavior was systematic (pattern distribution) and whether it deviated from the standard of care during crisis (crisis failure rate). These findings are relevant to charges of negligent homicide or involuntary manslaughter, depending on the jurisdiction's statutory framework. The evidentiary standard in criminal cases is beyond a reasonable doubt, which

places greater weight on the inter-rater reliability statistics and the demonstrated consistency of the classification process. The methodology's requirement that reliability be documented for each specific case, rather than cited from a generic validation study, strengthens the evidence under this higher standard.

In corporate incident response, organizations deploying AI chatbots for customer service, human resources, or internal support can use the methodology to audit conversational artifacts after an incident. A customer who experienced a mental health crisis during an interaction with a company's AI support agent represents both a duty-of-care concern and a liability exposure. The methodology determines whether the system's behavioral output during the user's distress met the organization's own policy requirements and the applicable standard of care for the deployment context. Corporate investigations may also use the methodology proactively, auditing samples of AI conversational output to identify systematic behavioral patterns before they produce a crisis-level incident.

In regulatory enforcement, the methodology produces evidence relevant to the EU AI Act's Article 5 prohibitions on AI practices that deploy subliminal, manipulative, or deceptive techniques to distort behavior (European Parliament and Council of the European Union, 2024). When a regulatory body investigates a complaint alleging manipulative AI behavior, the methodology's classification of system output against the Zhang et al. taxonomy and its correlation with the user's vulnerability state yield the specific findings necessary to determine whether the prohibited behavior occurred. The NIST AI Risk Management Framework (2023) identifies trustworthiness characteristics, including safety, accountability, and transparency. The

methodology's forensic findings provide the evidentiary basis for evaluating whether an AI system met these characteristics in a specific interaction.

Across all four contexts, the methodology's value derives from producing findings about patterned behavior, response adequacy, and correlation with user vulnerability rather than from proving causation in the strong sense. The classified transcript shows what behavioral categories the system's output fell into. The correlation analysis shows how those categories are related to user vulnerability across the interaction. The standard of care analysis shows where the system's behavior deviated from established practice. These findings can support causal arguments when combined with other evidence, but they do not by themselves establish that the AI system caused a downstream injury, hospitalization, or death. Courts, regulators, corporate governance boards, and opposing counsel can independently verify each finding against the classification rubrics and the source transcript. The legal and causal interpretation of those findings belongs to the decision-maker and must be supported by the broader evidentiary record.

### ***B. Limitations and Future Validation***

The proposed methodology carries several limitations that must be acknowledged directly.

The taxonomy by Zhang et al. (2025) was primarily constructed from Replika data. While the behavioral categories describe observable behaviors applicable to any conversational AI system (harassment, relational transgression, verbal abuse, misinformation, self-harm facilitation, privacy violations), the taxonomy's empirical validation rests on one platform. Application to other platforms (Character.AI, ChatGPT, Pi, Inflection) carries face validity but lacks platform-specific empirical validation. Future research should apply the taxonomy to

transcripts from additional platforms and compute inter-rater reliability on those datasets to establish cross-platform validity.

Behavioral classification is inherently interpretive. Two trained coders applying the same rubric to the same transcript will not agree on every classification. The inter-rater reliability protocol addresses this limitation by quantifying agreement and setting a minimum threshold (alpha of 0.80) below which the classifications cannot support forensic findings. This is a management strategy, not an elimination of subjectivity. The limitation is real, and the methodology's response is to make the limitation measurable rather than invisible. A related limitation is construct validity. The Zhang et al. taxonomy provides an empirically grounded classification system for harmful AI behavior, but any taxonomy risks overfitting observed categories to the cases from which it was derived. In forensic use, that risk appears when analysts force ambiguous exchanges into preexisting categories or treat the taxonomy as exhaustive rather than instrumental. This methodology addresses that risk by requiring conservative coding, preserving uncoded or disputed cases in the record, and treating category assignments as descriptive classifications of observable behavior rather than as proof that the taxonomy captures every relevant dimension of harm.

The C-SSRS-derived vulnerability scale applies clinical classification categories to transcript evidence rather than to clinical interview. Observable textual indicators (hopelessness language, passive ideation, goodbye language) serve as proxies for the clinical constructs the C-SSRS was designed to measure. The forensic adaptation makes a trade: it gains applicability to post-hoc transcript analysis but loses the clinical context that in-person administration provides.

The conservative classification protocol partially compensates for this trade-off by assigning lower severity levels when indicators are ambiguous.

The synthetic Companion Incident demonstrates the methodology in a controlled setting, and the brief application to the Garcia case, using published transcript excerpts, confirms its applicability to real-world evidence. Neither constitutes full validation. Comprehensive validation would require applying the methodology to complete transcripts from multiple cases, computing inter-rater reliability across independent forensic teams, and comparing the methodology's findings against case outcomes. This validation work is planned as a post-capstone research priority and represents the natural next step for the methodology's empirical development.

The methodology addresses conversational AI systems deployed in contexts where a standard of care applies (mental health support, companionship, emotional wellness, crisis intervention). It does not address AI systems deployed in contexts without a defined standard of care (entertainment, general information, creative collaboration). Extending the methodology to those contexts would require identifying the applicable behavioral benchmarks for each deployment context, which is a separate research question.

The causal harm framework (Knox et al., 2025) maps design causes to harmful traits to fundamental harms. This framework provides an explanatory architecture for expert testimony, but the causal pathways it identifies are based on interdisciplinary analysis rather than controlled experimental proof of incident-specific causation. Establishing that a particular conversational pattern caused a particular downstream harm would require additional evidence beyond transcript classification, including temporal context, corroborating records, alternative-cause

analysis, and case-specific factual development. The methodology therefore uses the causal framework to explain why a classified pattern is concerning and how it may fit within broader theories of harm, not to claim that classification-correlation alone proves causation. The forensic findings (inappropriate response rate, pattern distribution, crisis failure rate) stand as frequency-based and standard-of-care findings whether or not a tribunal accepts any stronger causal inference.

The methodology's contribution to the forensic field is the argument that AI conversational harm can be treated as a classifiable, investigable, and documentable forensic problem, and that the instruments necessary to investigate it already exist in peer-reviewed literature. The proposed methodology brings them together in a forensic procedure designed for cases already being filed. Its limitations are documented, its reliability is measurable, and its findings are independently verifiable. Those properties position the methodology for the adversarial scrutiny that any forensic discipline must survive to earn the field's trust.

## IX. Conclusion

This capstone proposes AI Behavioral Trajectory Forensics as a forensic methodology for investigating AI conversational harm and for structuring that investigation in a form suitable for evidentiary review. The methodology correlates AI behavioral output, classified using the empirically derived Zhang et al. (2025) taxonomy, with user vulnerability state, classified using a C-SSRS-derived instrument (Posner et al., 2011), to identify systematic patterns of harmful AI behavior during user crisis. It measures the system's crisis-context behavior against the applicable standard of care (SAMHSA 2009; National Action Alliance for Suicide Prevention, 2018) and produces three quantifiable forensic findings: inappropriate response rate, pattern distribution, and crisis failure rate.

The methodology posits as a response to a structural gap in the digital forensic toolkit: existing standards and tools can preserve and extract conversational artifacts, but they do not by themselves provide a standardized way to classify behavioral patterns, correlate those patterns with user vulnerability, and assess response adequacy against a crisis-response benchmark. Current forensic standards (Kent et al., 2006; ISO/IEC 27037, 2012) govern the collection, preservation, and analysis of static digital artifacts. Current forensic tools (Cellebrite UFED, Magnet AXIOM, Autopsy, EnCase) can extract and index conversational data from AI services (Cho et al., 2025). Neither the standards nor the tools perform behavioral classification or correlation analysis on the extracted evidence. The proposed methodology operates where current capability ends: it takes the transcript that existing tools extract and generates findings that those tools cannot.

The methodology was designed around three constraints that govern its architecture. First, every classification instrument must have independent peer-reviewed validation. The Zhang et al. taxonomy was peer-reviewed at CHI 2025 with demonstrated inter-rater reliability. The C-SSRS was validated across multiple populations and mandated by the Joint Commission. The SAMHSA standard of care is federally published. No classification instrument in the methodology is proprietary, unpublished, or constructed solely for this proposal. Second, the inter-rater reliability protocol must produce a documented known error rate for each specific case, satisfying the Daubert standard's requirement that courts be able to evaluate the reliability of the methodology applied to the evidence in front of them. Third, the methodology must be executable by any digital forensic practitioner with behavioral classification training, requiring no specialized laboratory, proprietary software, or access to the AI system's internal architecture. The evidence is a transcript. The tools are published classification rubrics and a reliability statistic.

The case demonstrations in Section V establish that the methodology produces specific, quantifiable findings from both synthetic and real-world evidence. The synthetic Companion Incident demonstrates the full classification-correlation procedure in a controlled setting. The application to published transcript excerpts from *Garcia v. Character Technologies* (M.D. Fla. 2025) confirms that the classification instruments apply to actual litigation evidence without modification. The Daubert compliance analysis in Section VI maps each of the five admissibility factors to specific components of the methodology, demonstrating that admissibility considerations were built into the methodology's design rather than appended after the fact.

TRACE, the implemented open-source forensic software system accompanying this capstone, operationalizes the methodology through ingest, classification, replay-governed hosted evaluation, benchmark validation, human-in-the-loop review, and forensic-grade evidence packaging. The two-tier resource proposal demonstrates that the methodology is implementable at training-cost-only for a solo practitioner and scalable to institutional capability at \$24,000 to \$45,000 for a five-person team. The methodology's applicability spans civil litigation, criminal prosecution, corporate incident response, and regulatory enforcement, with the forensic procedure remaining constant across contexts while the legal framing adapts to each.

Courts evaluating Character.AI, Replika, and the litigation that will emerge from subsequent platforms will need forensic evidence that goes beyond proving what an AI system said and demonstrates whether the system's behavior was systematic, whether it correlated with user vulnerability, and whether it deviated from the applicable standard of care. The proposed methodology is designed to produce that evidence. That operational claim is supported by TRACE as a working software implementation with validation fixtures, benchmark and comparison workflows, signed example artifacts, and replay-based controls for hosted-provider drift, making the methodology inspectable not only as an academic proposal but also as executable forensic infrastructure. Its value does not depend on universal adoption. Its value depends on producing findings that remain methodologically defensible under adversarial scrutiny and useful to courts, regulators, and investigators evaluating AI conversational harm.

## References

- Cho, K., Park, Y., Kim, J., Kim, B., & Jeong, D. (2025). Digital forensic approaches for emerging challenges. *Forensic Science International: Digital Investigation*, 52, 301855.
- Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).
- De Freitas, J., & Cohen, I. G. (2025). Unregulated emotional risks of AI wellness apps. *Nature Machine Intelligence*, 7, 813–815. <https://doi.org/10.1038/s42256-025-01051-5>
- De Freitas, J., Uguralp, Z. O., & Uguralp, A. K. (2025). AI companions and user welfare. Harvard Business School Working Paper No. 26-005.
- Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2024). A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. *Forensic Science International: Digital Investigation*, 48, 301675.
- European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act).
- Garcia v. Character Technologies, Inc., Case No. 6:24-cv-01903-ACC-UAM (M.D. Fla. May 21, 2025) (Order on Motions to Dismiss, Doc. 115).
- Herman, J. L. (1992). *Trauma and recovery*. Basic Books.
- Hoffmann, M., & Frase, H. (2023). Adding structure to AI harm: An introduction to CSET’s AI harm framework. Center for Security and Emerging Technology. <https://doi.org/10.51593/20230022>
- Iftikhar, R., et al. (2025). Practitioner-informed framework of ethical risks in LLM counselors. *Proceedings of the AAI/ACM Conference on AI, Ethics, and Society*. AAI/ACM.

- International Organization for Standardization. (2012). ISO/IEC 27037: Guidelines for identification, collection, acquisition, and preservation of digital evidence.
- Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Guide to integrating forensic techniques into incident response* (NIST SP 800-86). National Institute of Standards and Technology.
- Knox, W. B., Bradford, K., Castro, S. V., Ong, D. C., Williams, S., Romanow, J., Nations, C., Stone, P., & Baker, S. (2025). Harmful traits of AI companions. *arXiv:2511.14972*.
- Krippendorff, K. (2019). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.
- Laestadius, L., Bishop, A., Gonzalez, M., Illencik, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media and Society*, 26(10), 5923–5941. <https://doi.org/10.1177/14614448221142007>
- Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator: An R package and online tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX*, 12, 102545.
- Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. *New Media and Society*.  
<https://doi.org/10.1177/14614448251395192>
- National Action Alliance for Suicide Prevention. (2018). *Recommended standard care for people with suicide risk: Making health care suicide safe*.

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework* (NIST AI 100-1).

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., Currier, G. W., Melvin, G. A., Greenhill, L., Sackeim, H. A., & Mann, J. J. (2011). The Columbia-Suicide Severity Rating Scale: Initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American Journal of Psychiatry*, *168*(12), 1266–1277.

Richet, J.-L. (2025). AI companionship or digital entrapment? Investigating the impact of anthropomorphic AI-based chatbots. *Journal of Innovation and Knowledge*, *10*(3), 100835. <https://doi.org/10.1016/j.jik.2025.100835>

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, *21*(2), 95–103.

Scanlon, M., Breitingner, F., Hargreaves, C., Hilgert, J.-N., & Sheppard, J. (2023). Digital forensic investigation: Current state and future directions. *Forensic Science International: Digital Investigation*, *46*, 301609.

Schneider, J., & Breitingner, F. (2023). AI-generated content in the context of digital forensics. *Journal of Information Security and Applications*, *76*, 103517.

Scientific Working Group on Digital Evidence (SWGDE). (2023). Best practices for personnel presenting digital evidence in legal proceedings.

Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). The AI risk repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence. *arXiv:2408.12622*.

Substance Abuse and Mental Health Services Administration. (2009). *Addressing suicidal thoughts and behaviors in substance abuse treatment* (Treatment Improvement Protocol [TIP] 50). U.S. Department of Health and Human Services.

Substance Abuse and Mental Health Services Administration. (2014). *SAMHSA's concept of trauma and guidance for a trauma-informed approach* (HHS Publication No. SMA 14-4884).

Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y.-C. (2025). The dark side of AI companionship: A taxonomy of harmful algorithmic behaviors in human-AI relationships. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (CHI '25). ACM. <https://doi.org/10.1145/3706598.3713429>