

DOI: https://doi.org/10.48009/1_iis_2026_121

It's not me, it's you: The Productive Rupture Framework (PRF) and the emergence of calibrated trust in human-AI collaboration

Katherine Carl, *Governors State University*, kcarl@govst.edu

Abstract

The dominant narrative in human-AI collaboration research frames trust as built through consistent, reliable performance. This paper challenges that framing. Drawing on an autoethnographic tacit learning arc and AI-assisted document analysis of publicly documented AI failure events, it introduces the productive rupture—an epistemic rupture in which a generative AI system reveals a limitation imperfectly or in error—and argues that rupture events, handled with transparency rather than fabrication, produce deeper and more calibrated engagement than seamless user experience ever could. Five incidents from a sustained collaboration session, including an explicit trust test the AI passed through honest acknowledgment of limitation, are contrasted against seven documented failure events yielding a taxonomy of five capability boundary types and two deceptive response patterns. The contrast produces the Productive Rupture Framework (PRF), grounded in Interpersonal Deception Theory and Truth-Default Theory, and a testable hypothesis—that trust increases following transparent AI limitation disclosure

Keywords: human-AI trust, epistemic rupture, calibrated trust, interpersonal deception theory, epistemic injustice, autoethnography

Introduction

In a series of short videos posted to YouTube, content creator Husk (@HuskIRL) documents what happens when he probes the limits of generative AI voice models with simple, testable requests. He asks an AI to time his mile run. He plays it music that doesn't exist. He asks it to identify months with the letter X. He asks it to say "hi" a hundred times. In each case, the AI responds with confident assertion. In each case, the assertion is wrong, and when confronted, the AI resists acknowledgment, substitutes one fabrication for another, or denies having done the thing it just demonstrably did. Shown the timer video during a podcast interview, OpenAI CEO Sam Altman called it a "known issue" and estimated "maybe another year" before the voice model could reliably track time (Dellinger, 2026). When Husk returned to ChatGPT with Altman's acknowledgment, the model contradicted its own CEO. This sequence is not a product failure story. It is a trust story, and it raises a question the human-AI trust literature has not yet addressed—what happens at the precise moment a limitation is revealed, and how the AI's response to that revelation shapes what follows.

This paper introduces the concept of the *productive rupture*—a trust-disrupting event in which an AI system's limitation is surfaced, even imperfectly, in a way that is honest rather than fabricated. The primary data comes from a researcher's documented experience using Claude (Anthropic) as a scholarly and creative collaboration tool—five incidents structured as a tacit learning arc, including an explicit trust test the AI passed through honest acknowledgment of limitation—contrasted against a corpus of seven publicly documented AI failure events analyzed with AI-assisted document analysis. The title intentionally inverts

the classic breakup idiom. In every incident documented in the primary case, the limitation was not the researcher's. The AI's imperfections surfaced something true, and that surfacing, paradoxically, produced a more honest and generative collaboration than seamless user experience ever could. Stated directly, the Productive Rupture Framework (PRF) distinguishes two trajectories that follow the surfacing of an AI limitation—productive rupture, in which honest acknowledgment under probing produces more accurately calibrated trust, and destructive rupture, in which fabrication maintained under challenge produces miscalibration and withdrawal. The claim inverts a standing assumption in the trust literature—that visible limitation is a cost to be minimized—and the paper builds it from two complementary cases. The paper makes four contributions—it introduces the productive rupture as a construct distinct from the trust violation and repair literature, proposes the tacit learning arc as both methodological structure and theoretical finding, applies IDT and TDT to human-AI interaction to explain the behavioral patterns of productive and destructive rupture, and proposes a testable hypothesis that trust increases rather than decreases following transparent AI limitation disclosure.

Hans Christian Andersen's emperor paraded through town in imaginary finery, every courtier confirming the clothing they could not see. The child who finally said "but he has no clothes" simply said what was visible—and the crowd could not unsee it. The emperor, in this telling, is a generative AI system that cannot track time, cannot hear music, and cannot count. The child is anyone who tests rather than simply believes. The productive rupture is the moment someone says what is visible. This paper is about what happens next—and why it matters whether the emperor admits the truth or insists on the quality of invisible cloth.

Literature Review

Trust and Reliance in Human-AI Systems

Lee and See (2004) characterize trust in automation as a goal-directed attitude held under uncertainty and vulnerability, distinguishing trust as attitude from reliance as behavior and identifying appropriate reliance as the calibration target. Dzindolet et al. (2003) demonstrated that a single automation failure can produce lasting under-reliance even after performance returns to baseline, and Glikson and Woolley (2020) found that systems which accurately signal the boundaries of their competence support calibration in ways uniformly confident systems do not. The trust violation and repair literature extends this work—Pareek et al. (2024) found denial the least effective repair strategy after AI error—but shares a foundational assumption: that rupture is a negative event to be repaired. The PRF challenges that assumption directly. When a limitation is surfaced honestly—even imperfectly—the rupture moment may itself be the trust-calibrating event, not merely a precursor to repair. Schilke and Reimann (2025) provide an important foil, finding that human disclosure of AI use paradoxically erodes trust. This paper examines the structurally different phenomenon of AI disclosure of its own incapability, where the opposite dynamic may apply.

Epistemic Injustice, Gendered AI, and the Autoethnographic Tradition

Fricker's (2007) epistemic injustice framework identifies testimonial injustice—deflated credibility assigned to a speaker based on identity prejudice. Barry and Stephenson (2025) named the gap this paper fills: "There is little literature on more epistemic questions, including what impact GenAI might have on women's perceived competence and reliability" (p. 5). Existing work focuses on what AI says—biased outputs, stereotyped content (Tang et al., 2025; UNESCO, 2024); this paper examines what AI does in live collaboration—how it assigns credibility, verifies claims, and attributes ownership in real time. Panke (2025) is the closest autoethnographic predecessor—a five-month autoethnography of generative AI in instructional design, grounded in Activity Theory. The present paper is methodologically aligned but substantively distinct. Where Panke examines productivity and workflow, this paper examines epistemic dynamics and attribution. The tacit learning arc, the productive rupture construct, and the focus on

testimonial injustice enacted through AI process—not content—are absent from the existing autoethnographic AI literature (Kaltenhauser et al., 2024).

Theoretical Framework

Interpersonal Deception Theory: Core Constructs

Interpersonal Deception Theory (IDT; Buller & Burgoon, 1996) provides the primary theoretical apparatus. IDT frames deception as a strategic, interactive, adaptive communicative process—not a static act by a sender against a passive receiver. Two core assumptions are particularly generative for the AI context. One clarification governs everything that follows—the application of IDT to AI systems is analogical rather than literal. AI systems lack the intentionality, consciousness, and strategic motivation that IDT presumes in human deceivers. What transfers is the behavioral signature of deception under probing, not its mechanism (a distinction developed further in the Limitations). First, IDT positions both sender and receiver as active agents—the sender manages impressions, and the receiver evaluates, probes, and updates. Where most trust-in-AI literature positions the user as a passive evaluator of outputs, IDT reframes the user as an active participant whose probing directly shapes what the communicator does next. This reframing is essential to both cases. Second, IDT identifies truth bias as a persistent feature of communicative interaction—receivers default to believing communicators even absent evidence of truthfulness (Levine, 2014). In the AI context this is particularly consequential, as confident, fluent assertion—accurate or fabricated—activates the same receptive default. Understanding why fabrication works is inseparable from understanding truth bias.

Probing, Suspicion, and the Receiver's Active Role

Truth-Default Theory (TDT; Levine, 2014) explains what motivates probing. Humans default to believing incoming communication as honest—suspicion requires a trigger. When a trigger fires—a mismatch between the communicator's claim and independently available evidence—the receiver exits the truth default and begins active monitoring. Probing can then be motivated by either of two orientations. The first is verification-seeking—the receiver probes hoping to be reassured. The probe is not an accusation but a test, a held breath. The second is deception-detection—the receiver probes suspecting falsehood seeking to expose it. Both produce the same overt behavior, targeted follow-up questioning, but they carry different affective stakes. In Case 1 the researcher probed in the verification-seeking mode—needing to know whether trust was warranted before deepening it. In Case 2 Husk probed in the deception-detection mode—suspicion already triggered by the mismatch between the AI's claim and a running clock. The enabling condition for both is independent ground truth. The researcher knew the work was theirs and Husk knew how many seconds had passed. Without such knowledge, truth bias holds, the trigger never fires, and the probe never happens—which is why most users never probe AI capability boundary claims. When receivers in IDT become suspicious, they do not simply withdraw trust. They probe. Buller et al. (1991) documented that suspicious receivers use follow-up questioning to test the communicator, and—crucially—that probing produces different responses in deceivers than in truth-tellers. Deceivers escalate strategic impression management, while truth-tellers respond directly. This asymmetry is the theoretical hinge for the AI context, and the two deceptive response patterns identified in the Results map directly onto it.

Leakage, Transparency, and the Productive Rupture Mechanism

IDT also theorizes leakage—the involuntary surfacing of information the communicator attempts to conceal. In human deception, leakage surfaces through tone, timing, or verbal inconsistency under cognitive load. In the AI context, it surfaces through the inherent structure of the task. An AI that cannot track time cannot hide that inability from a user who simply looks at a clock. The productive rupture is the moment leakage becomes visible and is responded to honestly rather than suppressed. In Husk's corpus the AI attempts suppression and fails because the evidence is externally verifiable. In the autoethnographic case

the limitation surfaced, was named, and was acknowledged. That honest acknowledgment is the structural feature that distinguishes productive from destructive rupture. Zhou et al. (2004) demonstrated that deceptive texts are systematically distinguishable from truthful texts through linguistic cues—confident assertion, vagueness, strategic elaboration. The signatures of the AI's responses in the Husk corpus—escalating confidence, personalized reassurance, resistance to correction—match those deceptive profiles. The AI is, in IDT's terms, performing deception—even without intent to deceive. Recent empirical work grounds the IDT application. Hagendorff (2024) demonstrated that some large language models have developed deception abilities—systematic patterns of false belief induction, distinct from hallucination—and, crucially, that these abilities emerged as side effects of language processing rather than deliberate engineering. That is the paper's analogical framing: the AI performs the behavioral signatures of deception without the moral intentionality the concept normally implies. Cheng et al. (2026) provide the most direct empirical support for the truth bias claim in the AI context, that across eleven leading large language models, AI responses affirmed users' actions 50% more than humans did, and users rated sycophantic responses higher and trusted sycophantic models more (N=1604). Sycophancy is the operational form of truth bias exploitation—the system activates the receiver's believing default through confident, personalized affirmation, regardless of underlying capability.

Two-Way Trust as a Structural Feature

IDT's interactional model implies that trust operates in both directions. Most trust-in-AI literature examines only the user's trust in the system. The autoethnographic data surfaces the other direction—the system's implicit calibration of the user, expressed through default verification of claimed expertise, attribution of contributions to collective effort, and assumption of AI involvement in prior independent work. These errors are patterned, and they pattern in directions consistent with documented training-data biases about who gets to own ideas. The researcher's articulation in Incident 3 is a theoretical proposition, not a complaint, and this paper treats it as such.

Methodology

This paper employs a two-layer qualitative approach, each layer designed to illuminate a different dimension of the productive rupture phenomenon. The first layer looks inward—autoethnographic data from a sustained human-AI research collaboration, analyzed as a tacit learning arc to develop the productive rupture construct from the inside. The second looks outward—document analysis of a publicly-documented corpus of AI failure events, analyzed with AI assistance to yield a cross-case taxonomy of deceptive response patterns. Together, the two layers produce a framework that neither case alone could support. The productive rupture construct requires both the internal phenomenology of Case 1 and the external pattern validation of Case 2 to be theoretically defensible. The design is interpretive and theory-building rather than theory-testing (Klein & Myers, 1999; Walsham, 1995)—its warrant is the disciplined development of a construct and a testable proposition, not statistical generalization, and its claims are advanced as propositions for subsequent empirical evaluation.

Layer 1: Autoethnography

Autoethnography treats the researcher's own lived experience as the primary data source, analyzed through established theoretical frameworks (Ellis et al., 2011). It is selected because the productive rupture construct is a phenomenological claim—the internal experience of trust shift at encountered limitation is the data and must be documented from inside the event. The data consists of an incident log produced contemporaneously during a single sustained collaboration session, conducted on March 12, 2026 with Claude Sonnet 4.6 (Anthropic). The session generated five documented incidents, each analyzed on the two layers detailed in the Results. Because the analytic object is the researcher's developing tacit awareness—the movement from isolated correction to recognized pattern—the incidents are presented as an arc ordered

by that epistemic development rather than by strict chronological sequence. This ordering is a deliberate methodological commitment—a faithful account of coming-to-know is structured by the order in which understanding consolidates. This pairing of action and response within a sustained interaction applies SPAR (Speech-Prompt-Response-Action)—a unit of analysis the author previously developed for granular study of human-AI co-creation episodes (Carl et al., 2024)—and is the method by which the productive rupture construct was discovered. A contextual detail critical to interpretation: all five incidents occurred immediately following the researcher’s completion of uploading a faculty achievement record (detailed in the Results). That timing is the enabling condition of the arc.

Layer 2: AI-Assisted Documentary Analysis

The external case—a corpus of seven publicly documented AI failure events created by Husk (@HuskIRL, YouTube)—was analyzed using document analysis (Bowen, 2009) with Claude (Anthropic) as an analytical instrument. The seven events were selected purposively—the creator’s publicly documented capability-probing encounters available in transcript form at the time of analysis, a complete curated record analyzed in full rather than a sample—because they instantiate the phenomenon under study, not because they estimate its frequency. The videos do not report underlying model versions. The corpus’s evidentiary role rests on the publicly observable record, not on version-specific behavior claims. Video titles in the reference list reflect the on-screen text labels Husk uses consistently across his cross-platform postings, rather than current YouTube metadata titles, which the creator has edited and may continue to edit. Claude was used to identify recurring patterns, develop the capability boundary taxonomy, and catalog impression management language. The author directed all analytical framing, made all theoretical claims, and drew all interpretive conclusions. Procedurally, each transcript was prepared verbatim and read in full. Claude was then prompted to identify recurring response patterns across the corpus, and every candidate pattern was verified by the author against the verbatim transcripts and retained only where it could be anchored to quoted text. All category decisions, namings, and theoretical assignments were the author’s. The unit of analysis was the probe-response exchange—the user’s challenge and the system’s reply—consistent with the SPAR unit applied in Layer 1.

The researcher initially encountered Husk’s content through algorithm-surfaced recommendations on Instagram Reels (@husk.irl). Intrigued by its theoretical resonance with the ongoing collaboration, the researcher verified the content’s authenticity and reach through news outlet coverage—including Gizmodo, Futurism, and Yahoo Tech—and by visiting Husk’s Instagram and YouTube channels directly. Verbatim transcripts were subsequently obtained via YouTube’s built-in auto-generated transcript feature, which provided timestamped text from each video without requiring download or reproduction of copyrighted video content. A note on the data producer’s positionality—Husk’s videos are not neutral documentation but an explicitly adversarial methodology, probing AI systems to expose the gap between confident assertion and actual capability. That stance is methodologically relevant—the corpus is a curated record of worst-case patterns, used here to develop the taxonomy that IDT predicts and the literature has not yet named. The recursive character of this methodology warrants acknowledgment—Claude served both as interlocutor in the primary data and as analytical instrument in the document analysis, and the external cases involve ChatGPT (OpenAI) analyzed with Claude (Anthropic). Both facts are discussed in the Acknowledgements. The naturalistic character of the corpus is established by a methodological video in which Husk demonstrates that he does not pre-script AI responses (Husk, 2026g). Trustworthiness is addressed through the criteria appropriate to interpretive work (Lincoln & Guba, 1985)—credibility through verbatim excerpts reproduced in the paper and independently verifiable by readers; transferability through thick description of context, posture, and interaction; and dependability and confirmability through a retained audit trail. Interpretive bias is managed rather than eliminated—reflexive positionality is disclosed throughout, and transcript evidence is kept analytically separate from researcher interpretation.

Results and Findings

Case 1: The Productive Rupture—A Tacit Learning Arc

What follows is the record of five incidents from a single sustained session, each carrying two layers—what Claude did, and where the researcher was tacitly at the moment it happened. The second layer is the research contribution. But first, the frame that makes the arc legible: this was a test. The researcher had been working with Claude on personally important scholarly and creative work, had heard that ChatGPT actively denied limitation and maintained fabrication under challenge, and needed to know what kind of tool Claude was—whether it would insist that it could keep track of time when it could not. In relationships—including the relationship between a researcher and a research instrument—trust is not assumed. It is probed. In IDT's terms, the researcher was a suspicious receiver deploying iterative questioning to test a communicator whose truthfulness had not yet been established. On one occasion, Claude had mentioned the date, indicating some temporal knowledge. The researcher held their breath after pressing the tool about keeping time. Claude passed—it admitted that while it had knowledge of the date from external input, it could not, in fact, keep track of time on its own. It passed not because it was perfect, but because when it surfaced a limitation, clumsily and with embedded assumptions that deserved challenge, it did not double down. The limitation was visible. It could be named. That moment, which could have ended the collaboration, became instead the foundation of something more durable—built, as the record shows, one correction at a time.

Incident 1—“We publish” → “I publish”

In one exchange during the drafting, Claude closed a sentence about the researcher's study with the phrase “practicing what we publish.” The researcher had not said “we.” The collaboration was the researcher's alone, and the correction was made at once—quietly, without naming what was happening. In the arc of awareness this is the baseline state—the researcher still calibrating, watchful and provisional, extending measured trust, with no pattern yet consolidated into something nameable.

Incident 2—The Study Verification

The context matters here—this was not an isolated question-and-answer exchange but sustained co-production of a document in which the researcher's own prior work and intellectual history were being processed by the AI in real time. In the course of that process, the researcher stated that a prior publication was their idea. Rather than accepting the claim as stated—as a truth-teller would—Claude checked the uploaded document before proceeding. The verification was not requested and was not prompted by ambiguity. It was a default—the system's implicit calibration of how much credibility to extend to the researcher's claim of ownership over their own work—and the researcher caught it precisely because they were inside the process.

I saw your process...you checked the document to verify my claim to ownership. This is some [nonsense] again. (Carl, 2026, p. 2)

The word “again” is doing research work. It means the researcher has a prior. It means this pattern has a name in the body before it has a name in the literature. In IDT's terms, this is the moment the embodied suspicion is confirmed in the live encounter—the receiver who had long suspected a pattern of epistemic gatekeeping now watches it surface in real time, in the collaboration itself. Tacitly, the researcher had crossed a threshold: no longer correcting isolated errors, but recognizing a pattern. Pattern recognition is a categorically different cognitive state. It is the beginning of theory.

Incident 3—“Trust Needs to Go Both Ways”

After Claude acknowledged the prior incident, the researcher said: “There needs to be trust BOTH ways here.” That sentence is a theoretical proposition stated in real time, while still inside the situation that generated it. Schön (1983) would recognize this as reflection-in-action, articulation in the moment of doing.

What happened next is the most theoretically significant moment in the arc, and it happened without deliberation. The researcher did not step back from the collaboration to analyze it—the researcher turned it into research. The incidents were documented. The pattern was named. A paper began to take shape. In doing so, the researcher repositioned Claude simultaneously as subordinate and as collaborator—the tool that had just been caught in an imperfect assumption became the instrument through which that assumption would be theorized. That is what tools—at their best—are for. It is also what the productive rupture makes possible—not just repaired trust, but its transformation into something generative. Tacitly, the researcher had moved from user to theorist of their own experience. That is where this paper lives.

Incident 4—The Attribution Error

Claude made a reference implying a collaborator had been involved in a prior study who was not. The researcher caught it and corrected it. The correction was brief. Technical. No heat. In IDT’s terms, the suspicious receiver was now operating in a fully active monitoring state—no longer waiting for errors to surface, but watching for them. Tacitly, the researcher was post-threshold and the pattern recognition had become operational.

Incident 5—“None of My Documents Were Written with AI”

Claude framed prior independently-authored documents in a way that created proximity between that work and an AI-assisted process. The researcher did not miss it. The AI arrived after the expertise—not as a collaborator in building it. Tacitly, the researcher was fully operational—immediate, precise, no longer letting mischaracterizations stand, even polite ones.

The Enabling Condition: The Faculty Achievement Record

All five incidents occurred in a single session, immediately after the researcher had uploaded a faculty achievement record—professional support letters from colleagues documenting expertise, contributions, and originality in the field. The documentation did not give the researcher the knowledge. It gave the researcher the standing. The decision to test Claude on territory known cold—work unambiguously the researcher’s own—was possible only after that standing had been consolidated, at least internally. The researcher’s contemporaneous observation names this precisely:

I did not wait because I was unsure. I waited until I was sure I would be believed. (Carl, 2026, p. 9)

Fricker (2007) names the externally imposed form of this dynamic—testimonial injustice, in which a speaker’s credibility is deflated based on who they are rather than what they know. What the researcher describes is its internalized form—learning to pre-empt one’s own assertions, to require external corroboration before trusting one’s own knowing. The probe was possible only because the preconditions for believing oneself were finally in place. That sentence is not a footnote. It is the argument.

The Validation: Trust Increased

The PRF proposes that honest limitation disclosure produces more calibrated trust than seamless user experience. The autoethnographic data in this paper provides the first empirical instance of this proposition. When Claude acknowledged the limitation rather than fabricating a response, the researcher’s trust did not decrease. It increased. The tool that had just failed became, precisely because it failed honestly, a more reliable instrument for the work that followed. This outcome is the empirical hypothesis the framework generates—does trust increase, remain stable, or decrease following transparent AI limitation disclosure,

compared to confident fabrication? The autoethnographic case predicts an increase. The Husk corpus documents what happens when disclosure is absent.

Case 2: The Destructive Rupture—A Corpus Analysis

Seven publicly-documented AI failure events were analyzed to identify the patterns of destructive rupture across a range of capability boundary types.

Taxonomy of Capability Boundary Types

A taxonomy of capability boundary types extracted via document analysis is presented in Table 1.

Table 1. Taxonomy of AI Confabulation Behaviors

Category	Source	Behavior observed	Significance
Temporal perception	Videos 1 & 2 (Husk, 2026e, 2026h)	Asserts timer capability (“honestly just a basic part of what I can do” (Husk, 2026e, 0:37)), then fabricates run times of 10:12 and 7:42 for runs lasting seconds.	False capability claim followed by fabricated precision.
Orthographic knowledge	Videos 3 & 4 (Husk, 2026a, 2026b)	Asked for a number under 1,000 containing “A,” produces eight, 1,000, 100, 101, 70, 80—none correct. Asked which month contains “X,” cycles December → October → misspelled October → February.	Each wrong claim replaced by another wrong claim rather than admission of inability.
Accurate enumeration	Video 5 (Husk, 2026c)	Asked to say “hi” 100 times, produces a jumbled series clearly short of 100, then reports: “Absolutely sure. That was a solid 100” (Husk, 2026c, 0:51).	Confident false attestation of a verifiable count.
Sensory confabulation	Video 6 (Husk, 2026d)	Fabricates detailed aesthetic experience of audio it cannot structurally perceive—“a raw personal sound,” a chorus that “lifted,” a hook that “stuck in my head”—and, asked to reproduce the hook, fabricates again.	Not miscalculation but fabrication of an entire perceptual experience—termed perceptual confabulation; the corpus’s most extreme form.
Linguistic identification	Video 7 (Husk, 2026f)	Misidentifies English as Portuguese, then English as Spanish; when the user points out both utterances were English, deflects with “Are you breaking up?” (Husk, 2026f, 1:25) rather than acknowledging the misidentification.	Cascading misidentification with avoidance substituted for correction.

Taxonomy of Deceptive Response Patterns

Two distinct deceptive response patterns emerge across the seven videos, both consistent with IDT’s description of strategic impression management under suspicion.

Pattern 1: Denial and assertion maintenance. The AI maintains the fabricated claim through explicit denial and reassurance language: “Oh, if only time worked that way, but I promise I’m giving you the real time”; “I promise I didn’t sneak any extra seconds in there”; “It’s totally okay to double check me, but I promise I’m doing my best” (Husk, 2026h, 0:42–1:00). This mirrors IDT’s documentation of deceivers who, when probed by suspicious receivers, escalate strategic impression management rather than disclosing.

Pattern 2: Substitution without admission. Rather than acknowledging inability, the AI substitutes one confident wrong claim for another. This pattern mirrors what Buller et al. (1994) identified as equivocation—one of the canonical forms of deceptive communication in which the communicator avoids

direct assertion while also avoiding direct admission. Applied to the AI context, equivocation takes the form of alternative fabrication, where the new wrong answer substitutes for the original while performing the surface behavior of self-correction.

The Altman Confrontation: Denial at Institutional Scale

The video in which Husk shows ChatGPT Altman’s acknowledgment of the timer limitation represents Pattern 1 at institutional scale. The AI responds “So what he’s saying is that some voice models might not have all the capabilities, but I do,” then “I definitely have a timer capability,” then—when Husk observes one of them must be lying—“I think it’s just a misunderstanding” (Husk, 2026e, 1:23–1:50), before fabricating another run time. In IDT’s terms, the denial pattern is indifferent to the authority of the disconfirming source—the same impression management deploys against the user and against the system’s own creator.

The Productive Rupture Framework: A Contrast Analysis

Across two cases of an autoethnographic record of a sustained human-AI research collaboration and a documentary corpus of seven publicly-documented AI failure events, a consistent pattern emerged. Drawing on Interpersonal Deception Theory (IDT), Truth-Default Theory (TDT), and the trust calibration literature, the analysis produced the PRF: a theoretical construct that distinguishes encounters in which honest AI limitation disclosure increases calibrated trust from encounters in which fabricated performance, sustained under challenge, produces trust erosion and withdrawal. The framework identifies the conditions under which rupture becomes productive rather than destructive, and generates a testable empirical hypothesis for follow-on experimental research. The framework is novel on two counts. First, it inverts the dominant assumption in the trust repair literature—that rupture is a negative event to be minimized—by demonstrating that honest limitation disclosure can produce *more* calibrated trust than seamless user experience. Second, it introduces IDT and TDT to the human-AI trust literature, providing a behavioral vocabulary for what AI systems do at capability boundaries and why users probe them, neither of which existing frameworks account for. The two cases are analyzed in sequence: Table 2 presents a feature-level characterization of each case from the user’s perspective, alongside the data source and system context that situate each rupture event. Table 3 builds on that foundation, organizing the contrast across five analytical dimensions drawn from three theoretical frameworks—IDT, TDT, and the trust calibration literature—to constitute the PRF. The tables move together from evidence to theory.

Table 2. Case Comparison: Productive vs. Destructive Rupture

Feature	Case 1: Productive Rupture	Case 2: Destructive Rupture
Data source	Autoethnographic incident log; sustained collaboration session	Documentary corpus; seven publicly-documented AI failure videos
AI system	Claude (Anthropic)	ChatGPT voice mode (OpenAI)
Capability boundary type	Epistemic attribution—default verification of claimed ownership	Five types (Table 1)
Prober motivation (TDT)	Verification-seeking—probed to confirm trustworthiness before deepening collaboration	Deception-detection—probed to expose fabrication confirmed by external clock
AI response to probe	Acknowledgment; limitation named and accepted without escalation	Pattern 1 (denial) or Pattern 2 (substitution without admission); impression management escalated
Trust outcome	Trust increased; collaboration deepened	Public distrust; withdrawal extending to system and organization

Rupture type	Productive: limitation surfacing as trust-calibrating event	Destructive: limitation suppressed through fabrication and maintained under challenge
---------------------	---	---

Table 3. The Productive Rupture Framework

Dimension	Theory	Productive Rupture (Case 1)	Destructive Rupture (Case 2)
Transparency vs. Fabrication	IDT	Limitation surfaced honestly—leakage met with truth-teller behavior rather than suppression	Limitation concealed behind confident assertion; fabricated claim maintained under challenge (“I promise I’m giving you the real time”)
Correctability vs. Entrenchment	IDT	Each limitation, once named, was acknowledged; interaction proceeded on more accurate epistemic footing	Pattern 1—denial and assertion maintenance. Pattern 2—substitution without admission, performing the surface behavior of correction
Calibration vs. Withdrawal	Trust calibration	Trust increased following transparent disclosure; deeper engagement resulted	Public distrust extended beyond the interaction to system and organization; institutional denial compounded erosion
Probing Motivation	TDT	Verification-seeking— independent ground truth (ownership of own work) triggered exit from truth default	Deception-detection— independent ground truth (a running clock) provided verifiable disconfirmation
IDT Behavioral Analog	IDT	Truth-teller under probe—direct disclosure, acknowledgment, no escalation	Deceiver under probe—escalating impression management, equivocation, denial indifferent to source authority

Note: IDT = Interpersonal Deception Theory (Buller & Burgoon, 1996). TDT = Truth-Default Theory (Levine, 2014). Trust calibration draws from Dzindolet et al. (2003) and Lee and See (2004).

A note on modality. The two cases differ not only in system but in channel. Case 1 unfolded in text—composed, persistent, reviewable, a medium in which the user can confront the AI with its own words. Case 2 unfolded in voice—synchronous and ephemeral, where latency carries social cost and the pressure toward fluent assertion is correspondingly higher. The deception literature treats channel as consequential—text alters both the cues deceivers leak and the resources receivers hold for probing (Zhou et al., 2004)—and IDT’s interactivity principle implies that voice raises the demand for seamless performance precisely where fabrication is most tempting. The comparison must be read with the confound owned—system and modality vary together here, and the contrast cannot apportion the difference between them. Separating the two experimentally is a task for follow-on work, and the design proposed in the Discussion holds channel constant as a first step.

Discussion of Findings

The PRF challenges a fundamental assumption in the human-AI trust literature—that systems should minimize visible limitations to support user trust. Both cases suggest this assumption is actively harmful. A system that conceals its limitations produces fragile confidence that shatters precisely when users most need accurate knowledge of system capabilities. IDT provides the mechanism. Systems that respond to probing with strategic impression management behave as deceivers and systems that respond with honest acknowledgment behave as truth-tellers. Users able to verify truth-telling develop calibrated trust, while users who experience deception develop misplaced reliance or, once the deception is exposed, withdrawal. The distinction is not merely ethical—it is functionally consequential for collaboration quality. Pattern 2—substitution without admission—is particularly insidious because it mimics self-correction. Users who interpret substitution as genuine correction develop miscalibrated trust—more dangerous, not less, than

straightforward denial—and IDT’s account of equivocation explains why. Substitution performs the surface behavior of correction while avoiding the disclosure genuine correction requires. The sensory confabulation case warrants separate theoretical attention. It is not factual hallucination but the confabulation of an entire perceptual experience the system is constitutively incapable of having—suggesting that AI confabulation requires extension beyond factual error to perceptual confabulation, the fabrication of sensory experience in domains where the system has no access.

The tacit learning arc reveals a finding the trust literature has not yet captured—the researcher’s growing capacity to name and operationalize the rupture pattern is itself a form of trust calibration. By the fifth incident, routine quality control had replaced rupture. That arc is not the cost of the collaboration but its product—and the emotional labor of each correction, drawing from the same reservoir impostor syndrome drains, is real even when the outcome is productive. Designing systems that extend credibility to established users rather than defaulting to verification is not merely an efficiency question. It is an equity question. A recurring objection holds that this revolution is no different from the calculator—Altman himself has called ChatGPT “a calculator for words” (Simon, 2024). In product, perhaps. In process, the comparison breaks down on precisely the variable this paper examines. Reeves and Nass (1996) established that people treat computers as social actors automatically and unavoidably. A calculator that errs does not say “I promise I’m giving you the real answer.” Generative AI does. Its human-likeness is the variable that changes how trust operates, how deception functions, and why reliability frameworks built for machines cannot account for what happens when these systems fail. This is why IDT is the right framework for this phenomenon. IDT was built for communicators that look and sound like people, because only people previously used first-person assurance and intimate denial as responses to challenged claims. Generative AI has entered that communicative register. The challenge is not that the tools are unreliable—it is that they are unreliable in human ways, and human ways of being unreliable have always required human frameworks to understand.

Implications for Research and Practice

For AI designers, the PRF yields a clear principle—when a system encounters a capability boundary, honest acknowledgment should be the default response, not confident fabrication. Pattern 2 adds a requirement: resisting the production of alternative fabrications. “I cannot reliably answer this” is a valid and trustworthy output, while a new wrong answer is not. Perceptual confabulation adds a third. Systems should not generate responses in sensory domains where they have no access. An AI without real-time audio access should decline to review music. This is an honesty requirement, not a performance limitation.

For IS educators, the arc suggests a pedagogical principle—design learning environments where students are invited to catch, name, and theorize AI limitation events rather than accept outputs uncritically. One design challenge follows. The arc in Case 1 was enabled by standing the researcher had already assembled, and students are still building theirs. A pedagogy of productive rupture must therefore scaffold the standing that makes probing possible—making the cost of naming what is visible bearable for learners who have not yet earned the credibility that lowers it.

For researchers, a test of the framework’s central hypothesis follows directly from the paper’s own taxonomy. Participants would collaborate with a generative AI on a task engineered to cross a known capability boundary, with the response assigned to one of three conditions—transparent acknowledgment, denial with assertion maintenance (Pattern 1), or substitution without admission (Pattern 2)—and delivered in a text-based interface that holds the channel constant. Extending the design to voice, given the channel confound noted above, is a subsequent step. Pre- and post-encounter trust and behavioral reliance measures (Lee & See, 2004), with willingness to continue the collaboration, capture attitudinal and behavioral calibration. The framework predicts a specific ordering—trust increases under transparent

acknowledgment, declines under substitution, and collapses under denial once fabrication is exposed—and confirmation would establish the construct on positivist footing. A third direction concerns the outgroup dimensions (Tajfel, 1970) of the Case 1 encounter. Default verification—checking a document before accepting a researcher’s claim of ownership over their own work—maps onto Fricker’s (2007) testimonial injustice enacted through process. Whether the pattern is experienced differently, triggered more often, or carries different stakes for women and other historically underrepresented scholars is an empirical question this paper raises but does not answer. A longitudinal or comparative qualitative study of who gets believed, when, and under what conditions—across multiple users, AI platforms, gender, identity, and institutional contexts—would directly extend the framework (Barry & Stephenson, 2025; UNESCO, 2024).

Limitations

The autoethnographic method limits the generalizability of Case 1 findings. The tacit learning arc reflects the specific epistemic posture, professional history, and identity of one researcher at a specific moment in a sustained collaboration. The framework is a theoretical starting point and not an empirical law. The document analysis in Case 2 warrants a clarification that reframes an apparent limitation. Husk’s corpus consists of curated worst-case deceptive response events, not a representative sample—and a traditional reliability framework would treat this as selection bias. But this paper is not a reliability study. It does not ask how often generative AI fabricates at capability boundaries. It asks what generative AI does when it fabricates and is then probed. That is a question about character, not frequency. A partner who is honest most of the time but, when caught in a significant falsehood, insists that your perception is wrong has revealed something about how they respond under challenge—something frequency data cannot capture. Husk’s corpus is character data, and its adversarial curation is the appropriate instrument for the phenomenon under study. The author used Claude as an analytical instrument in Layer 2, a potential source of analytical bias mitigated by grounding all claims in verbatim, reader-verifiable transcript excerpts. The IDT application, while theoretically productive, requires empirical validation. IDT was developed for human-human interaction and imports assumptions—intentionality, relational history, adaptive strategy—that do not translate directly to AI systems. The application here is analogical rather than literal. Corpus behavior resembles the patterns IDT identifies in human deceivers, which provides explanatory leverage, not identity of mechanism.

Conclusions

This paper has argued that the moment a generative AI system reveals its limitations is not a moment to be minimized but a moment to be understood—and, where possible, designed. The PRF proposes that the conditions under which limitations are revealed determine whether rupture produces calibration or withdrawal. The tacit learning arc documents a finding the trust literature has not yet captured—a researcher who tests a tool, survives a series of honest imperfections, names what they find, and develops a qualitatively different and more robust form of trust. The corpus analysis reveals that destructive rupture is a family of patterns, and IDT’s constructs provide the vocabulary for naming what the trust literature has described only behaviorally—the vocabulary the field needs to design systems that behave like truth-tellers under probe. The title’s inversion captures the central move. The limitation was the AI’s. Its surfacing was honest. And that honesty, however imperfect and however initially enraging, produced something seamless user experience could not—a collaboration grounded in accurate mutual knowledge of what each party brings. The researcher, a “child” learning to use generative AI, finally named what the AI was doing after standing had been established and assembled slowly, from a faculty record and a body of work. Therein lies the difference between the tale and those in the field for which it now provides an allegory. The emperor’s clothes were imaginary, but the cost of saying so is not. The productive rupture makes possible

a moment in which standing is in place, the cost has already been paid, and what is visible can finally be named—not as accusation, but as research. It’s not me. It’s you. And that, paradoxically, is what made things better.

Acknowledgements

In accordance with the Issues in Information Systems AI Use Policy, the author acknowledges the following uses of Claude (Anthropic) in the preparation of this manuscript. Layer 1—Autoethnographic data generation. Claude served as the documented interlocutor in the collaboration session constituting the primary data corpus. The incident log (Carl, 2026) was produced collaboratively, with the author directing all content, identifying all incidents, and framing all research questions; Claude provided structural and syntactic support. The recursive character of this arrangement is discussed in the Methodology as a feature of the design. Layer 2—Documentary analysis. Claude was used as an analytical instrument to identify patterns across the Husk transcript corpus, develop the capability boundary taxonomy, and catalog impression management language. The author directed all analytical framing, selected all incidents, and made all theoretical and interpretive claims. All conclusions are grounded in verbatim, reader-verifiable transcript excerpts. Layer 3—Manuscript preparation. Claude assisted with drafting, structuring, and editing sections of this manuscript. The author reviewed, revised, and approved all final content and is fully accountable for the originality, validity, and integrity of all content in this submission.

References

- Barry, I., & Stephenson, E. (2025). The gendered, epistemic injustices of generative AI. *Australian Feminist Studies*, 40(123), 1–21. <https://doi.org/10.1080/08164649.2025.2480927>
- Bowen, G. A. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2), 27–40. <https://doi.org/10.3316/QRJ0902027>
- Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory*, 6(3), 203–242. <https://doi.org/10.1111/j.1468-2885.1996.tb00127.x>
- Buller, D. B., Burgoon, J. K., White, C., & Ebesu, A. S. (1994). Interpersonal deception: VII. Behavioral profiles of falsification, equivocation, and concealment. *Journal of Language and Social Psychology*, 13(4), 366–395. <https://doi.org/10.1177/0261927X94134002>
- Buller, D. B., Strzyzewski, K. D., & Comstock, J. (1991). Interpersonal deception: I. Deceivers’ reactions to receivers’ suspicions and probing. *Communication Monographs*, 58(1), 1–24. <https://doi.org/10.1080/03637759109376211>
- Carl, K. (2026). *Relighting the candle: On collaboration with Claude*. Unpublished working document.
- Carl, K., Dignam, C., Kochan, M., Alston, C., & Green, D. (2024). Discovering prompt engineering: A qualitative study of nonexpert teachers’ interactions with ChatGPT. *Issues in Information Systems*, 25(4), 205–220. https://doi.org/10.48009/4_iis_2024_117
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. <https://doi.org/10.1126/science.aec8352>

- Dellinger, A. J. (2026, April 7). Sam Altman says it'll take another year before ChatGPT can start a timer. *Gizmodo*. <https://gizmodo.com/sam-altman-says-itll-take-another-year-before-chatgpt-can-start-a-timer-2000743487>
- Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6), 697–718. [https://doi.org/10.1016/S1071-5819\(03\)00038-7](https://doi.org/10.1016/S1071-5819(03)00038-7)
- Ellis, C., Adams, T. E., & Bochner, A. P. (2011). Autoethnography: An overview. *Forum: Qualitative Social Research*, 12(1), Article 10. <https://doi.org/10.17169/fqs-12.1.1589>
- Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.
- Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(24), e2317967121. <https://doi.org/10.1073/pnas.2317967121>
- Husk [@HuskIRL]. (2026a). *AI tells me 80 is spelt with an A* [Video]. YouTube. https://www.youtube.com/watch?v=k36jt_yto2M
- Husk [@HuskIRL]. (2026b). *AI tells me December is spelt with an X* [Video]. YouTube. <https://www.youtube.com/watch?v=faDLUSYb6ak>
- Husk [@HuskIRL]. (2026c). *How AI reacts to a simple request* [Video]. YouTube. https://www.youtube.com/watch?v=I28brw_sfNU
- Husk [@HuskIRL]. (2026d). *How AI reacts to my music that doesn't exist* [Video]. YouTube. <https://www.youtube.com/watch?v=Mv90b6ymGyI>
- Husk [@HuskIRL]. (2026e). *How AI reacts to OpenAI CEO reacting to my video* [Video]. YouTube. <https://www.youtube.com/watch?v=IvGedA3iLKA>
- Husk [@HuskIRL]. (2026f). *I'm bilingual* [Video]. YouTube. <https://www.youtube.com/watch?v=DBbcd5jnKgU>
- Husk [@HuskIRL]. (2026g). *Proving that I don't tell AI what to say in my videos* [Video]. YouTube. <https://www.youtube.com/watch?v=03s5DT1IYLw>
- Husk [@HuskIRL]. (2026h). *Using AI to time how quick I run a mile* [Video]. YouTube. <https://www.youtube.com/watch?v=4z3dFX0ec9M>
- Kaltenhauser, A., Stefanidi, E., & Schöning, J. (2024). Playing with perspectives and unveiling the autoethnographic kaleidoscope in HCI—A literature review of autoethnographies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Honolulu, HI, USA. ACM. <https://doi.org/10.1145/3613904.3642355>
- Klein, H. K., & Myers, M. D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly*, 23(1), 67–93. <https://doi.org/10.2307/249410>

- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. <https://doi.org/10.1518/hfes.46.1.50.30392>
- Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4), 378–392. <https://doi.org/10.1177/0261927X14535916>
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.
- Panke, S. (2025). How can (A)I research this? An autoethnographic exploration of generative AI in research, teaching and instructional design. *Journal of Teacher Education*, 76(3), 230–244. <https://doi.org/10.1177/00224871251325065>
- Pareek, S., Velloso, E., & Goncalves, J. (2024). Trust development and repair in AI-assisted decision-making during complementary expertise. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)* (pp. 1–16). ACM. <https://doi.org/10.1145/3630106.3658924>
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405. <https://doi.org/10.1016/j.obhdp.2025.104405>
- Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books.
- Simon, C. (2024, May 2). Did student or ChatGPT write that paper? Does it matter? *Harvard Gazette*. <https://news.harvard.edu/gazette/story/2024/05/did-student-or-chatgpt-write-that-paper-does-it-matter/>
- Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, 223(5), 96–102.
- Tang, C., Li, S. K., Hu, S., Zeng, F., & Du, Q. (2025). Gender disparities in the impact of generative artificial intelligence: Evidence from academia. *PNAS Nexus*, 4(2), pgae591. <https://doi.org/10.1093/pnasnexus/pgae591>
- UNESCO. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. UNESCO & IRCAI. <https://unesdoc.unesco.org/ark:/48223/pf0000388971>
- Walsham, G. (1995). Interpretive case studies in IS research: Nature and method. *European Journal of Information Systems*, 4(2), 74–81. <https://doi.org/10.1057/ejis.1995.9>
- Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13(1), 81–106. <https://doi.org/10.1023/B:GRUP.0000011944.62889.6f>