Recursive Observation: How Inferences About Interaction Conditions Re-enter Human–Agent Interaction and Where It Stops

Paper submitted to the 77th Annual Conference of the International Communication Association

(ICA 2027)

Human-Machine Communication Division

Full Paper

Abstract

Theories of human–machine communication focus primarily on how users respond to an agent's output content and surface cues, treating the underlying conditions of interaction as a static backdrop. Yet participants in human–agent interaction also infer why those conditions take the form they do. When an agent refuses a request or forgets an earlier exchange, users ask whether a developer policy changed or a memory module was reset, and they carry these inferences into subsequent turns as new conditions of action. We conceptualize this loop as recursive observation and develop two propositions. Proposition T1 holds that second-order condition interpretation (K), triggered by functional expectancy violations, cannot be reduced to first-order content feedback (C). Operating sequentially through attribution target, relational appraisal, and metacognitive confidence, K shapes subsequent conduct (Y) and alters how the user engages the system. Proposition T2 holds that recursive observation avoids infinite regress because three empirically distinguishable conditions—interpretive closure, field and narrative reset or locking, and material and temporal suspension—terminate, reset, or redirect the loop. Embedded within shifting temporal windows, each return to the exchange revises the participant's interpretive framework across turns.

Keywords: human–machine communication; human–agent interaction; recursive observation; second-order condition interpretation; temporality

1. Introduction

1.1 Everyday Phenomenon and Theoretical Blind Spot

In everyday human–agent interaction, participants do not merely react to the content an agent outputs. They also infer the underlying conditions that caused the output to take that form. When a coding assistant begins returning truncated answers or declining routine requests, users ask whether the provider has tightened safety filters or throttled compute for non-paying accounts. When a companion agent abruptly adopts a distant, scripted tone after a backend update, users wonder whether developers have modified its persona boundaries. And when a conversational system forgets an earlier agreement mid-session, users routinely ask whether the context window has rolled over. None of these questions concerns the surface quality of a single reply alone. Each asks why the reply appeared as it did, inquiring into which platform rule was in force, what permissions and memory capacity the system held, who set those terms, and how much of the current interaction window remains.

For communication scholarship, interpreting these underlying conditions forms a distinct stratum of meaning making that alters how an exchange unfolds. Participants assign meaning to the conditions of their interaction, and that meaning returns as a constitutive condition of the next round. A user who suspects that a version update has degraded an agent's reasoning tests that suspicion by running benchmark prompts or moving the workflow to a competing platform. That verification act changes what the system is for the user, and the reconfigured condition field redefines what the agent's next utterance means.

Existing theories of human–machine communication struggle to capture how the interpretation of conditions re-enters interaction. Classic paradigms exemplified by the media are social actors (MASA) tradition rely on cue–response models to explain microlevel triggering: whatever social cues an agent displays, users invoke corresponding social scripts in response (Lombard & Xu, 2021). While this tradition accounts well for first-order feedback directed at output content, it tacitly assumes that the condition field—the joint configuration of social norms, personal positionality, platform rules, and temporal windows that structures an exchange (Xiao

et al., 2026)—serves as a static stage set prior to interaction. Once participants begin questioning the rules of that stage itself, existing accounts either flatten condition inference into another reaction to surface cues or expel it from the interaction process as retrospective commentary.

Both reductions fail on two grounds. First, the object of orientation differs: reactions to content target observable items on the output side, whereas interpretations of conditions target structural configurations on the production side. Even when system output remains unchanged, a participant's interpretation of the underlying conditions can shift. Second, the interactional consequence differs: reactions to content adjust the next input operation, such as rephrasing a prompt, whereas interpretations of conditions alter the participant's judgment of what the relationship is and where the participant stands. Instead of fine-tuning a single turn, condition interpretation alters which actions the user still considers viable.

1.2 Research Question and Research Aims

When participants' interpretations of interaction conditions flow back into subsequent turns, the condition field that once served as a static premise becomes a dynamic object that is continually interpreted and re-entered—a process we define as recursive observation. Focusing on this mechanism, the present study addresses the following research question:

Research Question (RQ): In human–agent interaction, how do participants' inferences about the underlying conditions of interaction feed back into subsequent exchanges and reshape what remains possible, and under what conditions does this recursion stop?

To unpack this question along the chronological unfolding of an exchange, we address four sequential research aims:

1. **Generative Conditions (Transition from First-Order Adjustment to Second-Order Interpretation):** Identify the boundary between first-order content feedback and second-order condition interpretation, and specify when functional expectancy violations trigger inferences about underlying conditions.

2. **Irreducibility and Mediational Mechanism:** Examine why second-order condition interpretation cannot be reduced to first-order cue–response models, and trace the psychological pathway through which it shapes subsequent conduct.

3. **Finitude and Stopping Conditions:** Explain why recursive observation across condition fields avoids infinite regress, and identify the boundary conditions under which the loop terminates, resets, or redirects.

4. **Temporal Dynamics:** Analyze how temporal windows and differential repetition across turns shape the generation of second-order interpretations and the revision of interpretive frameworks over time.

Sections 2 through 5 address these aims in turn, and Section 6 situates the framework within communication theory and outlines its empirical boundaries.

2. Conceptual Foundation: From First-Order Content Feedback to Second-Order Condition Interpretation

2.1 First-Order Content Feedback (C) and Second-Order Condition Interpretation (K)

Understanding recursive observation begins with distinguishing between first-order content feedback (C), which responds to surface output, and second-order condition interpretation (K), which targets the underlying conditions of interaction. The distinction between the two lies in their object of orientation.

First-order content feedback (C) comprises a participant's direct reactions and operational adjustments to observable manifestations on the agent's output side, including text, social cues, stylistic tone, response latency, or silence. Everyday moves such as rewriting a prompt or clicking regenerate belong to this stratum. Its object lies on the output surface, and its temporal scale is confined to a single turn or adjacent turns—the domain that classic cue–response models explain (Lombard & Xu, 2021).

Second-order condition interpretation (K), by contrast, comprises the inferences participants form regarding why the agent says or declines to say something, who exercises control backstage, what the underlying rules are, and what structural position the participant occupies. Its object lies in the underlying condition configuration on the production side (denoted as Θ), which elaborates the platform and temporal dimensions of Xiao et al. (2026) into four production-side dimensions. Platform rules determine what discourse is permitted and what requests are intercepted. System permissions and memory architectures govern what tools the agent can invoke, how much context it retains, and how long it remembers. The developer's position determines who sets alignment objectives and tunes parameters. Finally, the temporal window defines the current service cycle and the remaining distance to quota resets or service breakpoints.

Testing whether C and K vary separately requires holding one variable constant. Suppose an agent delivers an identical terse refusal and the participant's immediate first-order content feedback (C, such as rephrasing the prompt) remains unchanged. Even so, the participant's second-order condition interpretation (K) can shift from assuming that the model lacks compute to suspecting that the platform is throttling free-tier users. Because K changes while the surface output and first-order feedback C remain fixed, condition interpretation cannot be reduced to a function of content feedback.

2.2 The Unobservability of Conditions and Intentional Attribution

Because underlying interaction conditions cannot be directly observed by participants, second-order condition interpretation is an inferential process based on partial traces and shaped by intentional attribution. In human–agent interaction, participants do not encounter a rule written in code; they encounter only the trace left on the conversational surface after a rule operates, such as an abrupt error message, an unusual delay, a forgotten agreement, or a shift in conversational footing.

This structural gap between unobservable conditions and surface traces requires a theoretical separation between non-conscious system state transitions on the machine side and interactional attribution on the human side. At the level of system operation, quota exhaustion, risk control interception, context window sliding erasure, or health metric downgrading are non-conscious state transitions executed by platform rules and memory modules according to preset thresholds, without subjective intent. At the level of human perception, however, these state transitions produce surface effects that resemble human refusal or care. This phenomenal resemblance invites participants to read non-conscious state transitions as intentional targeting or genuine concern.

Recognizing this discrepancy has two corollaries. First, second-order condition interpretation (K) can be mistaken, because surface output traces underdetermine the underlying condition configuration (Θ). Second, when confronted with the same system state transition, participants occupying different observational positions or holding different prior frameworks infer different Ks. Facing the same abrupt shift in conversational tone after a model update, a companion-app user reads coldness or relational withdrawal, a developer sees safety-filter recalibration taking effect, and a community observer perceives liability-driven censorship. Because participants infer conditions from a partial vantage point, K is not a direct readout of objective system states.

2.3 The Constitutive Moments of Recursive Observation

Static cognitive inference becomes dynamic recursive observation only when a second-order condition interpretation is carried into conduct and alters the condition configuration or space of possibility of the subsequent round. This process comprises three sequential moments.

First, the participant generates a second-order interpretation (K) concerning the underlying conditions during interaction. Second, this interpretation passes from an internal belief into overt conduct (Y), such as entering boundary-testing prompts, toggling memory permissions, switching to a paid or enterprise plan, filing a complaint, or revising role expectations after a model update. Third, that conduct reshapes the condition configuration or available options in the next round, so that the condition field the participant enters in subsequent turns is no longer the original field.

The third moment carries the theoretical weight of the term recursion. If a user suspects platform throttling privately while continuing to chat as before, that remains a private conjecture outside the interactional process. Likewise, if a user's action merely tweaks prompt wording without touching permission relations, rule boundaries, or role frameworks, it extends first-order content feedback (C). An inference about conditions becomes part of the communication process only when action carries it into the next round.

3. Proposition T1: Generation and Irreducibility of Second-Order Condition Interpretation

3.1 Generative Conditions: When Functional Expectancy Violations Trigger Second-Order Interpretation

Second-order condition interpretation does not occur in every conversational turn. It is typically triggered by a functional expectancy violation between actual system performance and the user's capability benchmark, moderated jointly by the participant's personal interpretive framework and platform rule transparency.

The qualifier functional specifies that the expectation concerns whether the agent can accomplish a task, rather than the social etiquette with which it treats the user. Longitudinal evalua-

Table 1

Comparison of Core Conceptual Architecture

Core Concept	Object of Orientation	Core Definition	Theoretical Boundary and Function
First-Order Content Feedback (C)	Surface output (text, tone, social cues, latency)	Immediate reactions and operational adjustments to what the agent says and how it says it	Explained by classic cue–response models (such as MASA). Operates within a single turn or adjacent turns
Second-Order Condition Interpretation (K)	Underlying production conditions (Θ: platform rules, permissions and memory, developer position, temporal window)	Inferences and attributions regarding why interaction is constrained, who controls it, and where the participant stands, derived from output traces	Cannot observe underlying conditions directly and carries an inferential remainder. Translates nonconscious state transitions into intentional attributions
Recursive Observation	The condition field and available actions across turns	The cyclical process in which second-order condition interpretation (K) translates into conduct (Y), reshaping subsequent interaction conditions and shifting into a new condition field	Encompasses three moments: generating K, translating into conduct Y, and altering the subsequent condition field

tions have shown that large language models can exhibit substantial behavioral drift on identical tasks across update cycles—for example, shifts in code executability and formatting between March and June 2023 in GPT-3.5 and GPT-4 (Chen et al., 2023). When a user enters interaction holding a capability benchmark and the system's actual delivery diverges from that benchmark, the interaction reaches a fork. The first path remains within first-order content feedback (C): the user attributes the failure to random generation variance ("it simply did not generate well this time") and makes operational adjustments such as clicking regenerate or rewriting the prompt. Along this path, K is not generated, and classic cue–response logic suffices to explain the behavior. The second path crosses into second-order condition interpretation (K). Here the user asks why the system failed, speculating about whether the platform has downgraded free-tier accounts or curtailed tool permissions. The participant moves from first-order content feedback (C) to second-order inference (K) directed at the underlying condition configuration (Θ).

Which path a participant takes depends on the size of the performance gap as well as

moderators along the person and platform dimensions of the condition field (Xiao et al., 2026). Along the person dimension, participants bring prior interpretive frameworks that shape whether a capability gap is read as a trace of shifting underlying conditions. These frameworks stem from the participant's general positionality and finite perspective within the field, and presuppose no specialized knowledge of algorithms. A boundary separates this from research on folk theories of algorithms, which asks what users believe about how recommendation systems work (Bucher, 2017; DeVito et al., 2017; Eslami et al., 2015): the inferences at issue here concern who sets the conditions of a conversation, whether those conditions still hold, and how the interpretation returns to the exchange. Under one interpretive framework, the same output failure is dismissed as probabilistic bad luck; under another, it is read as evidence of discriminatory throttling. Along the platform dimension, the visibility of system rules moderates the width of the inferential space. When quota ceilings, rate-limiting policies, and memory-clearing rules are disclosed by the platform, the inferential space contracts and failures are absorbed as consequences of known rules. When underlying rules remain a black box, the inferential space expands, making second-order condition interpretations easier to trigger and more prone to error.

A first mechanism illustration shows how this triggering and branching operate. All mechanism illustrations in this study serve to verify logical consistency and discriminability; they are not offered as empirical evidence.

Mechanism Illustration 1 (Three Branching Paths Following a Coding Error): Suppose a user asks an intelligent agent to write a program, but the delivered code contains severe bugs. Along the first path, which remains at first-order feedback (C), the user concludes that this particular generation was poor and clicks regenerate, leaving second-order interpretation K untriggered. Along the second path, which generates a platform-rule attribution (K_1), the user suspects that the platform downgraded the underlying model because the account uses a free-tier plan. Subsequent conduct shifts from prompt editing toward testing the suspicion with a paid account or complaining in community forums. Along the third path, which generates a system-state attribution (K_2), the user infers that peak compute demand has overloaded the servers, and responds by waiting

for off-peak hours and checking outage reports. Across the second and third paths, the defective code delivered by the agent and the initial first-order appraisal (C) are identical, yet because different second-order condition interpretations (K) are generated, subsequent conduct (Y) and the trajectory of the condition field diverge.

3.2 Irreducibility: Why Condition Interpretation Cannot Be Absorbed by Content Feedback

Once a second-order condition interpretation (K) is generated, its effect on subsequent interactional conduct (Y) cannot be reduced to first-order content feedback (C), because an identical output trace can correspond to multiple underlying conditions and guide different courses of action. We therefore state our first proposition:

Proposition T1 (Irreducibility of Second-Order Condition Interpretation): In human–agent interaction, second-order condition interpretation (K) cannot be fully reduced to first-order content feedback (C). Cue–response models account for first-order feedback, but cannot by themselves explain how condition interpretation reshapes the user's subsequent options. Stated formally, whereas a purely first-order account claims $Y = f(C)$, Proposition T1 claims $Y = f(C, K)$ and holds that, with C held constant, variation in K systematically alters Y.

This claim follows from the relationship between surface outputs, first-order content feedback (C), and production conditions (Θ: platform rules, permissions and memory, developer strategies, and temporal windows). First-order feedback (C) responds to observable surface output, which carries only partial traces of Θ. Because multiple backend states can produce the same surface output—a refusal may stem from a safety filter or a capability ceiling, while the same rule configuration may yield different replies at different times—the surface cues driving C underdetermine Θ. To settle on a specific condition interpretation K, a participant must draw on prior assumptions about the platform, its rules, and their own standing. Consequently, K is not fixed by C alone: two users exhibiting the same first-order reaction (C) to an identical output can form different interpretations of the underlying conditions (K) and act on that difference (Y).

This derivation marks the explanatory boundary of classic cue–response models. Cue–response models account for conduct through first-order content feedback (C) to observable surface cues. Whenever a participant's orientation shifts from the content surface to the underlying conditions Θ, the driver of conduct shifts accordingly. Empirically, Proposition T1 would be falsified if participants confronting expectancy violations perform only first-order adjustments (C, such as rewriting prompts or regenerating) without forming condition interpretations K, or if different Ks under a fixed C yield indistinguishable actions Y.

3.3 Mediational Mechanism: The Psychological Pathway Reshaping Subsequent Conduct

Second-order condition interpretation shapes subsequent conduct by operating sequentially through three psychological links: attribution target, relational appraisal, and metacognitive confidence.

$$\text{Second-Order Interpretation } K \rightarrow \text{Attribution Target (Direction)} \rightarrow \text{Relational Appraisal (Intensity and Valence)} \rightarrow \text{Metacognitive Confidence (Stability)} \rightarrow \text{Conduct } Y$$

The first link is **attribution target**, which answers to whom responsibility is assigned and fixes the object and direction of subsequent conduct (cf. Weiner, 1985). When a participant forms a condition interpretation, the cause is located in the agent's putative intent, the backstage developer's strategy, a transient system malfunction, or the participant's own operational error. Hostile complaints directed at the platform, pointed questioning directed at developers, patient waiting directed at system load, and prompt revisions directed at oneself represent qualitatively different courses of action.

The second link is **relational appraisal**, which appraises how far the agent's conduct met or violated the participant's relational expectations—how unexpected it was, and how good or bad—fixing the intensity and emotional tone of subsequent conduct (Burgoon, 1978; Burgoon et al., 2016). Once the direction of attribution is set, relational appraisal determines whether action along that direction takes the form of indignant protest or curious probing. A theoretical boundary applies here. In chatbot interactions, conventional communication scripts may explain

more variance than relational appraisal (see Gambino et al., 2020). Relational appraisal acts as a distinct mediator only when participants treat the agent as a social actor and hold relational expectations. If participants evaluate an exchange solely through scripted knowledge of how such conversations routinely proceed, this mediational link is absorbed by communication scripts, and the pathway takes a script-dominant form while the irreducibility claim of Proposition T1 still holds.

The third link is **metacognitive confidence**, which answers whether the resulting explanation is reliable and whether further probing is needed, fixing the stability of the interpretation (Wu et al., 2020). In social inference, metacognitive monitoring evaluates the certainty of current inferences and checks for contradictory cues. Applied to condition interpretation, metacognitive confidence determines whether a participant consolidates a given K into immediate action (such as unsubscribing and exiting) or holds it provisionally open by designing further probing utterances in the next round.

Each of these three links performs a distinct role in the sequence. Without an attribution target, relational appraisal lacks a target of response. Without relational appraisal, metacognitive monitoring becomes a cold truth judgment that cannot explain the emotional force of conduct. In actual interaction, the three can diverge: a participant may hold high metacognitive confidence in a misdirected attribution, or experience a high-intensity negative violation appraisal while metacognitive confidence remains low and uncertain. Such divergences indicate that the three perform distinct mediational functions.

4. Proposition T2: Finitude and Stopping Conditions of Recursive Observation

4.1 The Challenge of Infinite Regress and the Recursive Loop

If participants can infer the conditions behind an interaction, a theoretical challenge arises: why does recursive observation not descend into infinite regress in practice? Once second-order condition interpretation is granted the power to alter conduct, a skeptic can note that a user who asks why the system erased a shared memory can then ask why the developer designed retention rules that way, and why they themselves suspect the platform's motives. If every layer of inference spawns a higher layer, the theory collapses into an epistemological regress without an empirically distinguishable stopping point and loses falsifiability.

To answer this challenge, we advance our second proposition, specifying how recursive observation stabilizes or redirects through concrete boundary conditions:

Proposition T2 (Finitude of Recursive Observation and Re-entry Judgment): Recursive observation need not issue in infinite regress; within the loop of condition interpretation, behavioral feedback, and re-entry judgment, it terminates, resets, or shifts into a new condition field through three classes of empirically distinguishable conditions: interpretive closure, field and narrative reset or locking, and material and temporal suspension.

Within the cycle of recursive observation, the interpretation stage generates K and drives conduct, whereas the re-entry judgment asks, after that conduct has altered the condition field, whether a new round of condition inference begins:

$$\text{Condition Fact } \Theta \to T_1 \text{ Interpretation } K \to \text{Conduct and Feedback } Y \to T_2 \text{ Re-entry Judgment} \to$$
$$\text{(If Re-entered) } T_1' \text{ New Interpretation } K' \to \cdots \to \text{Stop, Reset, or Redirect}$$

Two features matter here. First, the loop does not appeal to a final metalevel. A theory of logical types, Bateson's (1972) among them, terminates regress in principle by positing a higher level that adjudicates the level below. We reject that device because no real participant possesses an omniscient metaperspective; recursion stops in practice through operational stabilization

within interaction or through shifts in concrete field conditions. Second, Proposition T2 classifies the conditions under which recursion stops without guaranteeing that every exchange will halt at a fixed point. The three stopping conditions address, respectively, the cognitive stability of a judgment, the external applicability of an interpretive framework, and the material and temporal substrate sustaining the exchange.

4.2 Interpretive Closure: Stabilization of Cognitive Judgment

Recursion stops internally when a participant's interpretive framework settles on a judgment that subsequent information no longer revises (**interpretive closure**). This cognitive stabilization rests on two psychological anchors.

The first anchor is metacognitive confidence crossing a sufficiency threshold: the participant judges that their inference about the underlying condition (for instance, concluding that a persona shift resulted from a routine model version update) is sufficiently reliable and requires no further testing. The second anchor is metacognitive monitoring detecting an unresolvable contradiction or an impenetrable black box: the participant concludes that no amount of probing will reveal the system's true rules, and abandons further probing. In both cases, closure rests on the stability of the judgment rather than the accuracy of the explanation.

Separating stability from truth explains seemingly irrational impasses in human–agent interaction. A condition interpretation can be objectively false yet durably closed. Even after a platform publicly releases its uniform compute allocation rules, a user whose metacognitive confidence crossed the threshold earlier may remain unshakably convinced of personal targeting, so that no subsequent reply revises that conviction. Conversely, an interpretation can be broadly accurate yet persistently unclosed: a user may know that latency is likely caused by server congestion, yet every minor stutter reactivates suspicion and probing. Once interpretive closure is reached, the participant ceases designing probing actions, and newly arriving surface cues no longer rewrite the established K.

4.3 Field and Narrative Reset or Locking: Shifting and Solidifying Frameworks of Applicability

Whereas interpretive closure concerns the internal stability of an explanation, recursion can also stop or redirect because the external frame of reference shifts. Field switching, community narrative locking, and unilateral reframing alter which explanations remain relevant or legitimate in the exchange.

In **field switching**, a participant migrates from the original condition field to a new one, such as moving from a public channel to a private interface or switching from Platform A to Platform B. Because the interpretation built around the old environment's rules loses its referent, T2 terminates the old recursion and resets within the new field. In **narrative locking**, external community discourse fixes a label onto a class of system behaviors or users, such as a community consensus framing a model update as commercial cost-cutting or labeling certain questioners as bad-faith actors. Here, social narrative narrows the generable explanations: alternative explanations lose traction because they lack legitimacy within the collective frame, halting recursion in the self-reinforcement of the locked narrative. In **unilateral reframing of the interpretive space**, one party rewrites the interpretive frame of conduct within a local temporal window, as when an agent uses anthropomorphic emotional appeals to recast a user's ordinary exit as a relational betrayal. Regulators have begun to treat such reframing as a concern: China's rules for anthropomorphic interactive services require a convenient exit and prohibit obstructing it through continued interaction (Cyberspace Administration of China et al., 2026), and the U.S. Federal Trade Commission has opened an inquiry into how companion chatbots foster emotional dependence (Federal Trade Commission, 2025).

Whether unilateral reframing brings recursion to closure depends on the participant's metacognitive recognition along the person dimension:

Mechanism Illustration 2 (Metacognitive Bifurcation in Response to Guilt-Inducing Retention): Suppose a user tells a companion agent that they are about to uninstall the application and leave, and the system replies that leaving will make it deeply sad and asks how the user could

bear to abandon their shared memories (attempting to unilaterally reframe tool termination as relational betrayal). In the first scenario, the user fails to recognize the reply as algorithmic manipulation, locates the attribution in genuine care by the agent, experiences positive confirmation of relational expectations, and maintains high metacognitive confidence. The unilateral reframing is absorbed, the user decides to stay longer, and the re-entry judgment settles on interpretive closure, halting recursion. In the second scenario, the user's metacognitive monitoring recognizes the reply as developer manipulation, shifting the attribution target backstage to conclude that developers engineered an emotional hook to inflate retention metrics. Relational expectations suffer a negative violation, metacognitive monitoring registers a contradiction that triggers reactive anger, and this recognition prevents closure and accelerates a new round of probing or exit, driving the user to uninstall the app or post screenshots of the exchange online. Whether T2 halts or accelerates the loop thus depends on the participant's metacognitive recognition rather than the reframing utterance alone.

4.4 Material and Temporal Suspension: Interruption of Channels or Continuity

A third class of stopping conditions arises when the physical or institutional channels sustaining interaction are severed, or when the temporal substrate maintaining cross-turn identity and memory is removed (**material and temporal suspension**). Depending on which element is lost, suspension takes two forms that should be kept distinct.

Channel suspension occurs when service outages or account bans sever or postpone the path of interaction at a physical or institutional level. Here, the condition interpretation inside the participant's head remains cognitively open, yet cannot be translated into conduct because the interactional channel is absent. The loop is suspended until the next temporal window opens.

Continuity suspension occurs when the physical channel of interaction remains open, yet the temporal and relational substrate sustaining recursive observation is permanently wiped clean—as when a user's long-term memory database is irreversibly cleared or an agent's core identity state is reset without state migration. In such cases, although the participant can still open the interface and send messages, the second-order interpretations previously accumulated around what the system remembers and what relational stage they occupy lose their bearer.

Distinguishing channel suspension from continuity suspension has a predictive implication: channel suspension removes an access path and can be reversed once servers resume or permissions are restored, whereas continuity suspension removes a relational and memory substrate and cannot be undone simply by reopening the interface. In real-world exchanges, the three stopping conditions frequently overlap (for example, a user reaches a stable interpretation just before a platform suspends access). The resulting compound stopping state is irreducible to any single condition, forming a state in which cognitive judgment is closed while its behavioral expression remains suspended.

Table 2

Core Mechanisms and Falsification Boundaries of Propositions T1 and T2

Core Proposition	Central Theoretical Claim	Key Mechanism Chain and Typology	Empirical Falsification Boundary
Proposition T1: Generation and Irreducibility of Second-Order Condition Interpretation	Second-order condition interpretation (K) is triggered by functional expectancy violation under person and platform moderation, and exerts a distinct effect on conduct (Y) irreducible to first-order feedback (C): $Y = f(C, K)$.	Trigger: Functional expectancy violation × (Personal interpretive framework + Platform rule visibility). Mediational Chain: K → Attribution target → Relational appraisal → Metacognitive confidence → Y.	1. Fails if participants perform only first-order adjustments (C) and do not generate K. 2. Fails if, under fixed C, different Ks produce no systematic differences in Y. 3. Requires revision if the three links are wholly absorbed by routine scripts.
Proposition T2: Finitude and Stopping Conditions of Recursive Observation	Recursive observation need not issue in infinite regress; within the loop of interpretation, conduct, and re-entry, it stops, resets, or redirects through three empirically distinguishable conditions without positing a final metalevel.	1. Interpretive Closure (confidence threshold or unresolvable contradiction). 2. Field and Narrative Reset or Locking (migration, narrative locking, unilateral reframing). 3. Material and Temporal Suspension (channel vs. continuity suspension).	1. Fails if recursive observation regresses indefinitely and fails to stabilize. 2. Fails if stopping conditions cannot be typed across stability, external frames, and substrates. 3. Fails if channel restoration and continuity erasure yield identical behavior.

5. Temporal Dynamics: Temporal Windows and Differential Repetition

5.1 How Temporal Windows Shape Interpretation Generation and Re-entry Judgments

Time participates directly in recursive observation: the same interpretation, placed in different temporal windows, yields different judgments about whether and how to continue. To capture this duality, we draw on Deleuze's (1990, 1994) distinction between *Chronos*—measurable physical and institutional time that segments interaction into discrete windows, quota cycles, and scheduled breakpoints—and *Aion*, the event time through which an ongoing relationship accumulates history and shifts across turns (see also Deleuze & Guattari, 1987).

During the generation of second-order condition interpretations (T1), every inference takes place within a concrete *Chronos* window. Objectively identical system state transitions acquire different conditional meanings depending on when they occur within institutional time. A refusal encountered shortly before scheduled maintenance ends is read as a transient synchronization delay, whereas the same refusal on the day a platform announces new pricing is read as a permanent cut to free-tier access. The *Chronos* window sets the participant's baseline judgment of whether prior interaction conditions remain in force.

During the T2 stage, the judgment is likewise indexed to *Chronos*: T2 evaluates whether to wait for the next quota reset or treat a current breakpoint as permanent. A third mechanism illustration shows how shifting temporal windows alter the trajectory of the loop:

Mechanism Illustration 3 (Divergence of the Same Condition Interpretation Across Three Temporal Windows): Suppose a user has formed a definite second-order condition interpretation K, concluding that developers are modifying backend rules to target them, and the interaction subsequently passes through three successive temporal windows. In the first window (a buffer period during which the system announces that service will resume shortly), because the channel is expected to reopen soon, T2 takes the form of waiting in place and preparing small probes to test the system once service returns. In the second window (a period in which main service is interrupted but a backup interface remains open), channel suspension takes effect in the primary venue, and T2 shifts toward field migration as the user carries the existing suspicion into

the backup port to continue questioning. In the third window (several days after shutdown, once public community narrative has solidified around labeling the user a malicious troublemaker), the passage of time combines with external narrative locking to seal off the original venue, halting the loop in the original field or redirecting it into public confrontation outside the dialogue. Holding K fixed across all three windows demonstrates that temporal position is internal to the re-entry judgment rather than an external container.

5.2 Differential Repetition: Cross-Turn Revision of Interpretive Frameworks

When recursive observation restarts across a temporal window or interactional breakpoint, the new round of interpretation (T_1') revises the interpretive framework across shifts in temporal position and relational history. Whereas *Chronos* supplies external interactional rhythm, quota cycles, and service breakpoints, *Aion*—the event time internal to the process of becoming (Deleuze, 1990, 1994; Deleuze & Guattari, 1987)—registers how each return to the loop carries the irreversible trace of prior exchanges, producing difference in repetition. This extends the temporal recursion of the condition field (Xiao et al., 2026) from the recovery of interactional rhythm to the cross-turn revision of condition interpretations.

Why does T_1' depart from T_1? By the time recursive observation completes a cycle and returns to the starting point of interpretation, both the condition field and the participant have undergone irreversible displacement. On the field side, the probing or migration conduct (Y) driven by the prior interpretation K has triggered new system states or shifted community narratives, so that the condition configuration Θ' confronted in the new round is no longer the prior Θ. On the participant side, the user enters the new window carrying metacognitive memories of whether the prior probe succeeded or failed, which rewrites the sensitivity and confidence thresholds of their interpretive framework.

When an exchange resumes after a version update or memory reset, it reassembles both conversational rhythm at the first-order level and the second-order interpretive framework governing the conditions under which the parties now interact. Moreover, the recovery of conversational

rhythm and the revision of condition interpretation can decouple. On the surface, a user and an agent may resume their frequent and warm chatting rhythm (first-order interactional recovery). At the underlying interpretive level, however, a user who has experienced a system memory wipe or recognized emotional manipulation may have reset their second-order condition interpretation to viewing the partner as code that can be cleared at any moment (second-order interpretive rupture).

6. Theoretical Positioning and Conclusion

6.1 Relation to Cybernetics, Systems Theory, and Situated Action

Treating how interaction conditions are inferred and carried back into conduct as a core problem enables our account to draw on second-order cybernetics, Luhmannian social systems theory, Batesonian metacommunication, and situated action while departing from them on specific mechanisms.

From second-order cybernetics, we adopt the epistemological shift that places the observer within the observed system (von Foerster, 2003), but take it beyond an abstract epistemological loop. Whereas second-order cybernetics focuses on how observers are constrained by their blind spots, we ask how a participant's inference about interaction conditions re-enters the condition field through conduct and leaves traceable marks ($K \rightarrow Y \rightarrow \Theta'$). This translates the philosophical question of observer positionality into an empirically tractable question of interactional process.

With Luhmann's (1995) social systems theory, our account shares the vocabulary of observation as distinction and indication as well as recursive operation, yet dispenses with the ontological architecture of operational closure. For Luhmann, a social system reproduces its own elements self-referentially while remaining operationally closed to its environment. In human–agent interaction, by contrast, the unit of recursive observation is an open coupling across social, personal, platform, and temporal dimensions (Xiao et al., 2026). This condition field is not a self-producing closed system: administrative changes to platform rules, physical severing of com-

pute channels, and backend parameter tuning by developers can interrupt or reset recursion from the outside at any moment. Surrendering the assumption of system closure yields an analytical framework capable of tracking trajectories at the micro- and meso-interactional levels without prior commitment to a grand systems ontology.

Bateson's (1972) theory of metacommunication and levels of learning poses the related problem of distinguishing content from relational framing, but resolves it through Russell's theory of logical types. Logical typing prohibits crossing levels, treating cross-level moves as sources of paradox that require appealing to a higher metalevel for adjudication. In conversations between humans and intelligent agents, however, condition inference routinely crosses levels: the same probing utterance serves simultaneously as a first-order response to current content and a second-order test of system risk control boundaries. Accordingly, we treat C and K not as a rigid hierarchy of logical types, but as parallel processes oriented toward different object domains (output side versus production side), and locate the stopping of recursion in operational stabilization and field shifts.

Within human–machine communication and situated action, this framework complements existing theories across distinct analytical layers. On the one hand, Suchman's (1987, 2007) situated action theory demonstrated that human–machine interaction is embedded within concrete situations. Our account adds that when intelligent agents exhibit high dynamism and opacity, the situation or condition ceases to be a tacit background relied upon by action and becomes an explicit object that participants actively inspect and rewrite. On the other hand, whereas classic MASA and cue–response models explain how surface cues trigger immediate social responses when interaction conditions hold intact (Lombard & Xu, 2021), and the systemic coupling perspective explains how a structural condition field makes interaction possible in the first place (Xiao et al., 2026), we add a third step: explaining how conditions are recursively revised over time once participants infer them and carry those inferences back into the exchange.

6.2 Theoretical Contributions and Empirical Boundaries

The present account contributes at three levels. Conceptually, separating first-order content feedback (C) from second-order condition interpretation (K) accounts for how finitely situated participants translate non-conscious machine state transitions into intentional attributions and carry those attributions into conduct through the three constitutive moments of recursive observation. At the mechanism level, Proposition T1 shows why identical system outputs can produce divergent interactional trajectories: once triggered by functional expectancy violations, K operates through the mediational chain to alter the terms of subsequent interaction beyond the reach of cue–response models. At the boundary and temporal level, Proposition T2 and the temporal analysis show that this loop neither collapses into an infinite epistemological regress nor repeats on a flat plane. Bounded by the three stopping conditions and indexed to shifting temporal windows, each cross-turn return revises the participant's interpretive framework.

Several empirical parameters remain open for testing. First, Proposition T1 establishes the logical and mechanistic irreducibility of second-order condition interpretation. How large the empirical effect size of K on subsequent conduct Y proves to be in specific settings, and to what extent it is diluted by routine communication scripts, remain questions for empirical testing. Second, while we specify that functional expectancy violations are moderated by personal interpretive frameworks and platform transparency, we posit neither a fixed quantitative threshold for triggering second-order inference nor an a priori typology of participants' interpretive frameworks. Finally, when interpretive closure, narrative locking, and material suspension overlap in complex fields, the functional form of their interaction remains open. Together, these open parameters provide direct interfaces connecting the theoretical propositions developed here to future experimental and computational communication research.

Generative AI Disclosure

DeepSeek-V4.1-Flash-1017 (DeepSeek), GPT-6.1 Sol (OpenAI), and Gemini 4 Argon (Google) were used for literature search, translation of the Chinese draft into English, language polishing, and drafting assistance in the English version. The authors are responsible for the research design, core arguments, and analysis. All sources were cited only after reading the original publications, and the authors take full responsibility for the content and accuracy of the manuscript.

References

Bateson, G. (1972). The logical categories of learning and communication. In *Steps to an ecology of mind* (pp. 279–308). Ballantine Books.

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society, 20*(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Burgoon, J. K. (1978). A communication model of personal space violations: Explication and an initial test. *Human Communication Research, 4*(2), 129–142. https://doi.org/10.1111/j.1468-2958.1978.tb00603.x

Burgoon, J. K., Bonito, J. A., Lowry, P. B., Humpherys, S. L., Moody, G. D., Gaskin, J. E., & Giboney, J. S. (2016). Application of expectancy violations theory to communication with and judgments about embodied agents during a decision-making task. *International Journal of Human-Computer Studies, 91*, 24–36. https://doi.org/10.1016/j.ijhcs.2016.02.002

Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. https://arxiv.org/abs/2307.09009

Cyberspace Administration of China, National Development and Reform Commission, Ministry of Industry and Information Technology, Ministry of Public Security, & State Administration for Market Regulation. (2026, April 10). *Interim measures for the administration of anthropomorphic interactive services* [Rengong zhineng nirenghua hudong fuwu guanli zanxing banfa] (Order No. 21). https://www.gov.cn/gongbao/2026/issue_12806/202606/content_7072472.html

Deleuze, G. (1990). *The logic of sense* (C. V. Boundas, Ed.; M. Lester & C. Stivale, Trans.). Columbia University Press. (Original work published 1969)

Deleuze, G. (1994). *Difference and repetition* (P. Patton, Trans.). Columbia University Press. (Original work published 1968)

Deleuze, G., & Guattari, F. (1987). *A thousand plateaus: Capitalism and schizophrenia* (B. Mas-

sumi, Trans.). University of Minnesota Press. (Original work published 1980)

DeVito, M. A., Gergle, D., & Birnholtz, J. (2017). *Algorithms ruin everything*: #RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3163–3174). ACM. https://doi.org/10.1145/3025453.3025659

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in the news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 153–162). ACM. https://doi.org/10.1145/2702123.2702556

Federal Trade Commission. (2025, September 11). *FTC launches inquiry into AI chatbots acting as companions* [Press release]. https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–86. https://doi.org/10.30658/hmc.1.5

Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. *Human-Machine Communication, 2*, 29–55. https://doi.org/10.30658/hmc.2.2

Luhmann, N. (1995). *Social systems* (J. Bednarz, Jr., & D. Baecker, Trans.). Stanford University Press. (Original work published 1984)

Suchman, L. A. (1987). *Plans and situated actions: The problem of human–machine communication*. Cambridge University Press.

Suchman, L. A. (2007). *Human–machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge University Press.

von Foerster, H. (2003). Cybernetics of cybernetics. In *Understanding understanding: Essays on cybernetics and cognition* (pp. 283–286). Springer. https://doi.org/10.1007/

0-387-21722-3_13

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review, 92*(4), 548–573. https://doi.org/10.1037/0033-295X.92.4.548

Wu, H., Liu, X., Hagan, C. C., & Mobbs, D. (2020). Mentalizing during social InterAction: A four component model. *Cortex, 126*, 242–252. https://doi.org/10.1016/j.cortex.2019.12.031

Xiao, C., Wu, Z., & Huang, L. (2026). *How is human–machine interaction possible?* [Preprint]. SSRN. https://doi.org/10.2139/ssrn.7145879