Kybernetes

Beyond Watzlawick: axioms of pragmatic interaction for second-order cybernetics

Yiannis Laouris
Futures Design Unit, Future Worlds Center, Nicosia, Cyprus and International Society for the Systems Sciences, Cookeville, Tennessee, USA

Received 30 October 2025
Revised 25 January 2026
11 March 2026
23 April 2026
Accepted 4 May 2026

Abstract

Purpose – To formalize Watzlawick, Beavin and Jackson's pragmatics of human communication as a level-separated, testable axiom system suitable for second-order cybernetics, so that paradox, reflexivity and context-dependence are treated as phenomena to be modeled rather than pathologies to be excluded.

Design/methodology/approach – We restate five classical propositions as axioms for unavoidability, meta-inevitability, code choice, observer-dependent punctuation and relational dynamics (A1–A5), and extend them with five interactional axioms (A6–A10) for turn-taking, repair, common ground, relevance defaults and public–private channel separation. We derive short theorems, specify counter-models to test axiom independence, and propose interpretive posterior entropy as an observable for code-variance and leakage effects.

Findings – The axiom set preserves Watzlawick–Bateson insights while making them empirically usable. It explains how divergent punctuation can produce reciprocal blame, how repeated meta-signals stabilize complementary or symmetric dyads and how code/leakage choices constrain interpretive uncertainty. The framework applies to human–human and socio-technical (human–AI, multi-agent) interaction.

Research limitations/implications – The paper is primarily theoretical; empirical validation requires interactional corpora annotated for relational/meta content, timing/repair moves and code/leakage features, plus experimental manipulations of observer segmentation. Future work can implement belief-update operators (e.g. AGM-style) and compare entropy-based predictions across media.

Practical implications – The axioms can inform the design of dialogue protocols, training for mediation/facilitation and human–AI interfaces that accommodate mis-punctuation and repair. Implementing A6–A8 in dialogue managers should reduce human–machine misunderstandings.

Social implications – By formalizing how escalation and relational locking emerge from observer-dependent segmentation, the framework helps diagnose and pre-empt interactional pathologies in organizational, intercultural and participatory/SDD settings.

Originality/value – The article shows that pragmatic interaction can be rendered in an observer-dependent, second-order cybernetic calculus that yields testable predictions — something often considered impossible for paradox-rich communication. It bridges Bateson–Watzlawick pragmatics, conversation analysis and contemporary systems/cybernetics.

Keywords Second-order cybernetics, Communication, Information theory, Systems theory, Social cybernetics, Mathematical modeling

Paper type Conceptual paper

Plain language summary

This paper explains, in simple terms, how everyday communication can be described with a small set of rules. We start from the well-known ideas of Paul Watzlawick and Gregory Bateson, who showed that people always communicate (even when they are silent), that people often interpret messages in relational terms ("what this says about us"), and that people often disagree about who "started" a problem. We turn these ideas into clear, testable statements (axioms) and show how they work together.

Kybernetes
Vol. 55 No. 13, 2026
pp. 69-103
Emerald Publishing Limited
e-ISSN: 1758-7883
p-ISSN: 0368-492X
DOI 10.1108/K-10-2025-2605

The paper also adds rules from conversation analysis, such as taking turns, repairing misunderstandings, building common ground and keeping some information private. Together, these rules can explain why two people looking at the same interaction can reach very different conclusions, and why some conversations get stuck in blame or dominance patterns.

Because the rules are written in a formal way, they can be used not only for human–human talk but also for human–AI systems. This is important today, because AI often misreads pauses, timing or indirect signals. Our framework shows designers where these misreadings happen and how to fix them. In short, the article offers a bridge between classic communication insights and modern socio-technical systems.

Introduction

Paul Watzlawick and colleagues famously articulated five propositions about human communication that have shaped interaction research and practice for decades (Watzlawick *et al*., 1967). Building in part on Bateson's ecology of mind, where information is a "difference that makes a difference" in systems of interacting observers (Bateson, 1972; Ruesch and Bateson, 1951), their account reframed everyday talk and silence alike as communicative acts embedded in relational patterns. In compact form, the five "axioms" (after Scott, 1997; Watzlawick *et al*., 1967) are:

A1. One cannot not communicate

Whenever behavior is observable (e.g. speech, gesture, even silence) it can change the other's beliefs; in our terms, any observable behavior $B\_i(t)$ can trigger a belief update in j: $Bel\_j(t)$ is j's beliefs at t (defined formally below).

A2. One cannot not metacommunicate

Every communicative uptake includes both content and a relationship attribution; the relationship aspect "classifies" the content, i.e. stance/role/affect are always being conveyed alongside semantics (a distinction anticipated in information theory by MacKay's, 1969, structural vs. semantic information, though without the relational focus we develop here).

A3. Communication can be analog(ical), digital, or mixed

Messages may use continuous/iconic cues (e.g. prosody, posture) "analog," discrete symbols (e.g. words, numbers), "digital," or both; code choice systematically changes how much interpretations vary across receivers.

A4. Communication is punctuated differently by sender and receiver

People slice the same stream into episodes at different cut points; these segmentation functions lead to divergent attributions about who initiated, escalated, or responded.

A5. Communication patterns are symmetrical or complementary

Dyads tend toward balance/equality vs. difference/hierarchy; repeated meta-signals update a relational state π_ij that stabilizes either symmetry (near zero) or complementarity (above a threshold).

These propositions, while framed informally, anticipated key insights in formal pragmatics regarding context-dependence, implicature, and the action-constituting nature of utterances (Levinson, 1983).

Despite their enduring influence, these are not axioms in the strict sense. First, they mix levels of analysis, i.e. information, semantics/pragmatics, cognitive attribution, and social dynamics, so logical independence is unclear. Second, several propositions are difficult to falsify or operationalize as stated (e.g. "one cannot not communicate" under what

observability and belief-update conditions?). Third, the set interleaves descriptive claims (what typically happens), normative hints (how coordination ought to proceed), and metaphoric distinctions (digital/analogic) without specifying measurement models or proof obligations. Finally, key constructs (e.g. "relationship aspect," "punctuation") are conceptually rich but underspecified for estimation, making it hard to derive non-trivial theorems or to test competing explanations (cf. Bateson, 1972).

We aim to go beyond Watzlawick not by discarding his insights, but by making the Watzlawick–Bateson pragmatics explicitly formal, level-separated, and empirically testable within a sociocybernetic frame.

We (1) sharpen definitions of agents, behaviors, message structure, segmentation (punctuation), and relational stance; (2) restate the core claims as level-separated axioms designed for mutual consistency and approximate independence; (3) derive theorems with explicit proof sketches (e.g. communicativity of silence; punctuation asymmetry yielding reciprocal blame; meta-signal accumulation stabilizing symmetry/complementarity); and (4) propose empirical procedures for coding relational content, estimating dyadic stance, and experimentally manipulating segmentation. The intended result is a sociocybernetic calculus that preserves the spirit of the original program while making it cumulative, refutable, and practically diagnostic. Our formalization demonstrates that axiomatization is possible even for reflexive, paradoxical phenomena traditionally viewed as resisting formal treatment, and that doing so enhances rather than diminishes their practical utility.

This move is especially relevant for second-order cybernetics, where observers and their distinctions are part of the system to be described (von Foerster, 1974, 1979). Similar ambitions can be seen in conversation-theoretic approaches to interaction (Pask, 1975, 1976) and in later second-order formulations that treat communication as an observer-dependent construction (Glanville, 2002, 2004). Our contribution is to show that this line of work can be rendered as a level-separated axiom system and linked to operational measures such as interpretive entropy (cf. Brier, 2008). What is novel here is not the general claim that interaction is observer-dependent, but the construction of a minimal, level-separated axiom system with independence tests, derived theorems, and operational observables for pragmatic interaction.

This contribution also speaks directly to current debates in Kybernetes on the contemporary scope of cybernetics and systems theory. Recent work in the journal has renewed attention to second-order operational epistemology (Roth and Sales, 2025; Saratxaga Arregi, 2025), to efforts to bridge living, technical and socio-technical systems within a more unified cybernetic ontology (Osejo-Bucheli, 2025), and to interaction across architecture, psychology and technology in emerging socio-technical settings (Temizel, 2025). Our paper enters this conversation by moving from general observer-dependence to a minimal, formal, level-separated axiomatics of pragmatic interaction: it specifies the minimal distinctions an observing system must track if communication is to be modeled, compared and tested across human–human and human–AI contexts.

Paradox, reflexivity, and context-dependence are not pathologies to be ruled out (as in classical logical-type approaches) but the very phenomena that a viable contemporary cybernetic theory must be able to formalize and test (von Foerster, 1974, 1979).

Why formalization has proven difficult

Despite more than five decades of influence, Watzlawick's five axioms have resisted systematic formalization. A review of the literature reveals not a crowded field of failed attempts, but rather a conspicuous absence: the axioms have been applied, critiqued, and extended clinically, yet rarely subjected to the mathematical rigor their name suggests they deserve. We identify three substantive attempts and patterns that reveal why formalization remains elusive.

(1) Bateson's deliberate abstention (1951–1972). As documented by Lutterer (2007), Watzlawick's axioms derive almost entirely from Gregory Bateson's prior work on

cybernetic communication, schismogenesis, and logical typing. Yet Bateson himself "abstains from a formal connection of his insights. He delivers no theory as a whole" (Lutterer, 2007, p. 1025). Bateson's Steps to an Ecology of Mind (Bateson, 1972) is rich with conceptual machinery, e.g. difference, information as "news of difference," levels of learning, but presents no axiom system, no proof obligations, and no falsification criteria. His resistance was philosophical: he viewed formal systems as potentially reifying what should remain fluid, pattern-based thinking. What was missed: By avoiding formalization, Bateson left key constructs (e.g. "logical types," "context") operationally undefined, making empirical tests impossible. Watzlawick inherited this richness but also its vagueness.

(2) Watzlawick *et al.*'s own hedged formalism (Watzlawick *et al.*, 1967). The original Pragmatics of Human Communication presented the axioms as "tentative" and explicitly disclaimed completeness (Watzlawick *et al.*, 1967, chapter 2 title, but also 2.7 Summary, p. 70). While the book invokes mathematical terminology, e.g. "calculus," "logical types," "set theory," etc., these references are metaphorical rather than operational. The authors provide no formal semantics, no independence proofs, and no derivation rules. As del Rio (2012, p. 343) notes in his retrospective review, "it remains unclear how they came upon these premises," and the formalization "culminates" in the five axioms but proceeds no further into theorems or measurement models. What was missed: Without explicitly specifying scope conditions, belief-update operators, or relational dynamics, the axioms remained heuristics, powerful for reframing clinical cases, but not constraining enough to rule out alternative explanations or generate novel predictions beyond intuition.

(3) Systemic therapy's pragmatic turn away from formalism (1974–2007). Following Pragmatics, Watzlawick's subsequent work (Watzlawick *et al.*, 1974; Watzlawick, 1984) moved decisively toward radical constructivism and therapeutic intervention techniques, abandoning the project of formal theory-building. The Brief Therapy Center at the Mental Research Institute in Palo Alto prioritized what works clinically over what can be proven formally. Meanwhile, followers diverged: Haley (1963, 1987) and the strategic therapy school operationalized interventions without formalizing principles, while Luhmann's (1995) social systems theory adopted cybernetic language but built an entirely separate formal apparatus. What was missed: The clinical success of Palo Alto methods created little incentive to formalize. Practitioners could apply "one cannot not communicate" or "punctuation" metaphorically in session without needing to specify Bel_*j*(*t*) or P_*i* functions (*vide infra*). The axioms thus became therapeutic folklore, evidence-informed, yes, but not evidence-constraining.

(4) Bavelas's empirical disconfirmation (2025; *vide infra*). In a striking recent development, Janet Beavin Bavelas, original co-author, now states that microanalysis of face-to-face dialogue shows "some [axioms] have worked pretty well, others do not" and that the sender-receiver model underlying them has been "scientifically refuted" (Godat and Czerny, 2021). Her video-based conversation research reveals communication as faster and more co-constructed than Watzlawick's framework assumed. However, Bavelas has not proposed a formal alternative; instead, she calls for abandoning the axiom framework entirely in favor of moment-by-moment empirical description. What was missed: While identifying failures, Bavelas offers no salvage operation; no specification of which axioms fail under what conditions, no formal counter-model, and no replacement calculus.

The pattern across attempts: Formalization has been avoided (Bateson), gestured at but not executed (Watzlawick *et al.*), or abandoned in favor of practice (systemic therapy) or rejected

empirically without formal reconstruction (Bavelas). The core obstacle appears to be that Watzlawick's most profound insights (i.e. paradox, reflexivity, context-dependence) are precisely the phenomena that resist standard formal treatment. This obstacle is now especially salient in the AI era, where human–AI and multi-agent systems routinely generate meta-communication, mis-punctuation, and emergent relational dynamics that demand explicit, testable formalisms rather than ad-hoc heuristics. Whitehead and Russell (1910–1913) identified them as the chief culprit of problems in logic, considered them an abnormality of language, and proposed a 'theory of logical types' to rule them out of the language of science. Watzlawick and Bateson insisted paradox is constitutive of human interaction. To formalize Watzlawick is thus to formalize what formal systems classically prohibit. Second-order cybernetics, however, requires precisely this: formalisms that can include the observer, recursion, and meta-communication.

Our contribution is to accept this challenge directly: we retain paradox and reflexivity as first-class phenomena (e.g. via T1-T3 on blame and relational lock-in, and T6 on symmetrical escalation, and T7 on meta-communication paradoxes; *vide infra*) while providing the formal scaffolding that makes them tractable for measurement and intervention.

Methods

Minimal formal framework

The framework below names the objects we track in interaction (agents, behaviors, interpretations, relations) so later axioms can be stated without ambiguity and mapped to measurements.

To turn the classic propositions into testable science, we use a lean ontology that makes the moving parts of interaction explicit. Our design principles are *parsimony* (as few primitives as possible), *level-separation* (information, coding, cognition, and social dynamics kept distinct), and *operationalizability* (every construct must admit observation, estimation, or manipulation). By fixing these primitives, we can restate Watzlawick's insights as axioms with clear independence claims and falsifiable consequences, paving the way for proofs, metrics, and interventions.

Agents and time. Agents $A = \{1, \ldots, n\}$ interact in discrete ticks $t = 1, 2, \ldots$

Behaviors. Each agent i produces an observable behavior $B_i(t)$ (*speech, gesture, gaze, silence/withdrawal* $\varnothing$, etc.).

Observation and interpretation. If agent j can observe $B_i(t)$, they form an interpretation $M_{i\to j}(t) = I_j(B_i(t), \textit{context})$.

Non-ontological stance. We treat observable conducts $B_i(t)$ as action schemata that *afford* readings but do not contain meaning. Meaning (including relational meaning) arises only when an observer j applies an interpretive operation I_j within a punctuated context p_j, yielding a situated reading $M_{i\to j}(t)$. Thus, conducts do not "carry" a relational level; rather, relational inferences are attributed by observers under conventions and roles. This aligns with second-order/constructivist views where communication is constituted by selections of observers situated in interaction.

Message structure. Each message M has:

(1) Content C(M): task/semantic proposition(s).

(2) Relational/meta R(M): stance/role/affect/rights–duty cues.

This content/relational bifurcation echoes MacKay's (1969) foundational distinction between structural information (what choices were made from available options, roughly our C(M)) and semantic information (what those choices mean for action and response, roughly our R (M)). MacKay observed that the same structural signal could carry different semantic import depending on receiver state—anticipating our observer-dependent interpretation functions I_j and the inevitability of relational uptake (A2). While MacKay's framework remained

primarily dyadic and information-theoretic, our formalization extends it by making relational dynamics (A5) and punctuation-dependent attribution (A4) explicit operational commitments.

Codes and channels.

- Digital = discrete/symbolic (words, numbers).
- Analog = graded/iconic (prosody, posture, distance).
- Encodings may be mixed across channels (voice, text, gaze, timing . . .).

Punctuation (segmentation). Each agent has a segmentation function p_i that slices the event stream into episodes and assigns causal/response roles.

Relational stance. For dyad (i, j), let $\pi\_ij(t) \in \mathbb{R}$ encode relative dominance/affiliation (near 0 = symmetric; $|\pi\_ij|$ large → complementary).

Estimating $\pi\_ij(t)$. Empirically, $\pi\_ij(t)$ is treated as a latent ordinal state inferred from meta-signals (e.g. interruptions, compliance bids, mitigation, honorifics, repair acceptance). A simple implementation is a state-space model (e.g. Ordinal Dynamic Generalized Linear Models, GLMs, and Hidden Markov Models, HMMs) in which these surface cues are observations and $\pi\_ij(t)$ is the hidden trajectory. This yields a time series for $\pi\_ij(t)$ compatible with A5's update rule.

Belief update criterion (communication as information). Let $\mathrm{Bel\_}j(t)$ be j's beliefs at t.

We say i has communicated to j at t if $\mathrm{Bel\_}j(t+1) \neq \mathrm{Bel\_}j(t)$ in part due to observing $\mathrm{B\_}i(t)$.

Belief states and update dynamics. We model each agent j's beliefs at time t as a set of graded commitments:

$$\mathrm{Bel\_j(t)} = \{(\mathrm{p\_1}, \mathrm{c\_1}), (\mathrm{p\_2}, \mathrm{c\_2}), \ldots, (\mathrm{p\_m}, \mathrm{c\_m})\}$$

where p_i is a proposition (e.g. "agent i is cooperative," "the request is face-threatening") and $c\_i \in [0,1]$ is j's confidence in that proposition. We say j believes p at t if $(p, c) \in \mathrm{Bel\_}j(t)$ with $\mathrm{c} > \theta$ for some threshold θ (typically $\theta = 0.5$, adjustable by context).

Update Operator. The function $\mathrm{U\_}j : \mathrm{Bel} \times \mathrm{Behavior} \times \mathrm{Context} \rightarrow \mathrm{Bel}$ maps the prior belief state, observed behavior, and contextual features to a posterior state. Here, Bel is the space of belief states, Behavior the space of observable behaviors, and Context task/environmental features used by $\mathrm{I\_}j$.

Formally:

$$\mathrm{U\_}j\big(\mathrm{Bel\_}j(t), \mathrm{B\_}i(t), \mathit{context}\big) = \mathrm{Bel\_}j(t+1)$$

where $\mathrm{Bel\_}j(t+1)$ differs from $\mathrm{Bel\_}j(t)$ in at least one of three ways:

(1) *Confidence shift*: $(p, c) \in \mathrm{Bel\_}j(t)$ *and* $(p, c') \in \mathrm{Bel\_}j(t+1)$ *with* $c' \neq c$

(2) *Proposition addition*: $(p, c) \in \mathrm{Bel\_}j(t+1)$ *but* $(p, \cdot) \notin \mathrm{Bel\_}j(t)$

(3) *Proposition removal*: $(p, c) \in \mathrm{Bel\_}j(t)$ *with* $c > \theta$ *but* $(p, c') \in \mathrm{Bel\_}j(t+1)$ *with* $c' \leq \theta$

$\mathrm{U\_}j$ is a formal placeholder specifying that updates occur; empirical implementations must estimate it from data (e.g. via regression of confidence changes $\Delta\mathrm{c}$ on features of $\mathrm{B\_}i(t)$ and context, or Bayesian networks calibrated to participant response patterns).

Communication criterion (*A1 operationalized*). We say agent i has communicated to agent j at time t if observing $\mathrm{B\_}i(t)$ induces a non-trivial update: $\mathrm{Bel\_}j(t+1) \neq \mathrm{Bel\_}j(t)$ in the sense that at least one confidence score changes by more than ε (minimal detectable difference, e.g. $\varepsilon = 0.1$), or a new proposition enters above threshold θ.

Worked example: silence as communication (T1 preview)

Context: i and j are in a meeting; j asks i a direct question

Prior: $\mathrm{Bel\_}j(t) = \{$("i is cooperative", 0.8), ("i has information", 0.7)$\}$

Observation: B_$i(t)$ = ∅ (intentional silence, no response after 3 seconds)

Update: U_j interprets silence in context → Bel_$j(t+1)$ = {("*i is cooperative", 0.3*), ("*i has information"*, 0.5), ("*i is withholding"*, 0.7), ("*topic is face-threatening", 0.6*)}

Result: Multiple confidence shifts (cooperative down, new propositions added) ⇒ communication occurred per A1, even though B_$i(t)$ was "doing nothing."

This example illustrates that silence is a behavior value (∅ ∈ range(B_i)), and in co-presence with sufficient observability, it induces belief change—hence it communicates.

Derivation conventions

This section specifies the formal machinery we use in the rest of the paper. Readers who prefer the big picture can skim for the bolded definitions and the summaries; the Applications and Interventions section shows how each piece is used empirically. In brief: we define states (beliefs, relations), operators (update, segmentation), and observables (interpretive entropy) that let the axioms yield testable predictions.

This paper is theoretical. "Results" are theorems and corollaries derived from the axioms within a minimal formal framework; empirical procedures are proposed as mappings from the formal primitives to observable measures. This style of derivation follows the second-order cybernetics insistence that distinctions are made by an observer situated in the system (von Foerster, 1974; Glanville, 2002).

Scope Conditions. Unless otherwise stated, proofs and sketches assume:

(1) Co-presence/observability: when we assert "j observes B_$i(t)$," the event is publicly observable for the relevant agents (no hidden channels), with noise bounded below a stated threshold.

(2) Monotone belief update: Bel_$j(t+1)$ = U_j(Bel_$j(t)$, B_$i(t)$, *context*) is well-defined and (weakly) sensitive to new observations; if B_$i(t)$ is informative under context, then Bel_$j(t+1) \neq$ Bel_$j(t)$ (A1).

Definition. U_j denotes agent j's belief-update operator, mapping prior beliefs, the observed behavior B_$i(t)$, and context to a posterior belief state: U_j: B × O × X → B.

Belief updates can be implemented by AGM-style operators (Alchourrón *et al*., 1985).

(3) Level separation: information flow (A1), meta-content (A2), code/medium (A3), punctuation/attribution (A4), and relational dynamics (A5) are treated as distinct assumptions; additional axioms A6–A10 are add-ons that must not contradict A1–A5.

Primitive Objects and Notation. Agents $A = \{1, \ldots, n\}$; behaviors B_$i(t)$; interpretations/messages M_$\{i \to j\}(t)$ = I_j(B_$i(t)$, *context*); content C(M) and relational/meta R(M); segmentation functions p_i; dyadic relational state $\pi\_ij(t) \in \mathbb{R}$; belief state Bel_$j(t)$; timing norms Θ; repair moves ρ; encoding policy κ_i; leakage λ.

Indicator status. The variables we deploy (e.g. H_j(M), $p\_j(r|\mathrm{M})$, $\pi\_i\_j(t)$) are partial indicators reconstructed from interpretive traces under explicit coding and punctuation assumptions. They are not direct "measures of the relational field," but operational summaries that support comparative and falsifiable claims.

Relational Dynamics. When needed, we instantiate A5 with a simple state-space update

$$\pi\_ij(t+1) = \pi\_ij(t) + \mathrm{f}\big(R\big(\mathrm{M}\_\{i \to j\}(t)\big)\big) + \eta\_t,$$

where f is sign-preserving for dominance vs. leveling meta-signals and $\eta\_t \sim N(0, \sigma^2)$ i.i.d. with $\sigma^2 < K$ (bounded noise). Symmetry/complementarity is defined by a threshold ε: $|\pi\_ij(t)| \leq \varepsilon \Rightarrow$ symmetric; otherwise complementary.

Functional form. f : $R \to \mathbb{R}$ extracts signed dominance from relational cues. A minimal linear implementation is f(R(M)) = Σ_k w_k · r_k(M), where *r_k* are behavioral indicators (interruptions, volume, spatial positioning, compliance resistance, etc.) and w_k are context-specific weights estimated empirically. Mixed signals are aggregated linearly; non-linear extensions (thresholding, interaction terms) can be tested against data.

Interpretive Posterior and Entropy. Receiver-specific posterior over relational interpretations:

p_*j*(*r* |M) for *r* in *R* (*R* = set of relational categories), produced by I_*j*.

p_*j*(*r* |M) = probability (posterior) that receiver *j* assigns to relational interpretation *r* given message M. It's a distribution over categories in *R* (e.g. dominance, leveling, affiliation).

Posterior interpretive entropy (discrete R: Shannon, 1948; see also Cover and Thomas, 2006):

$$\mathrm{H\_}j(M) = -\mathrm{sum\_}\{r \text{ in } \mathrm{R}\}\,\mathrm{p\_}j\left(r \mid M\right) * \log \mathrm{p\_}j\left(\mathrm{r} \mid M\right)$$

(Logs are natural by default → units "nats"; base-2 allowed → "bits".)

Group summary (mean entropy across receivers):

$\mathrm{H\_bar}(M) = (1/\mathrm{N}) * \mathrm{sum\_}\{j = 1..\mathrm{N}\}\mathrm{H\_}j(M)$[1]

Continuous relational parameter (e.g. *π_ij):*

Differential entropy: h_*j*(*π* |M) = - ∫ *p_j*(*π* |M) * log *p_j*(*π* |M) d*π*.

Constructing *R*(M). In empirical applications, *R*(M) is instantiated from a coding scheme that lists admissible relational readings (e.g. leveling, dominance, deference, challenge, repair-initiation). The set is finite and predeclared for the corpus/experiment, derived from prior literature or pilot annotation, and may include an "other/indeterminate" catch-all. This makes the interpretation space reproducible across studies.

Estimating *p_j*(*r*|M). We estimate *p_j*(*r*|M) from interpretive traces using Bayesian logistic/softmax models with features of M (code, timing, leakage, punctuation). In all cases, *H_j*(M) summarizes the dispersion of readings conditional on the chosen R(M), features, and punctuation *p_j*.

Scope and limits. *H_j*(·) indexes the dispersion of possible readings after an interpretive operation *I_j* under a chosen punctuation *p_j*. It does not measure "meaning" itself nor the quality/viability of an interaction. Lower entropy can signal stronger constraints or tighter conventions, but it need not be better: rigid, face-threatening patterns may have low entropy yet poor relational outcomes. Accordingly, we use entropy for descriptive/comparative predictions, while evaluation relies on viability criteria (e.g. sustained coordination, repair success, facework adequacy).

Comparisons/language

"Entropy decreases/increases" = directional change in H_bar(M) between conditions while holding C(M) and context fixed.

Partial order across codes (A3): e.g. H_bar(M | mixed) ≤ H_bar(M | digital); we do not assert a total order for all pairs.

Requisite interpretive variety and concentration. Because lower entropy can indicate either efficient coordination or rigidity/dominance, we distinguish requisite from non-requisite variety by task/context. Alongside *H_j*(M), we report a concentration/asymmetry indicator (e.g. Gini or Herfindahl index over *p_j*(*r*|M); optionally KL-divergence to a context-specific target distribution). This separates "low *H*" due to healthy convergence from "low *H*" due to suppressive patterns, and prevents an implicit negentropic normativity.

Leakage effect (A10): H_bar(M |λ) follows an inverted-U with a minimum at λ* (too little or too much leakage raises uncertainty).

In applications and Table 2, entropy is used descriptively (as dispersion over a predefined interpretation set *R*(M)); it is not a proxy for relational quality. We further distinguish required vs. non-required interpretive variety for the task at hand.

Notation. We use Cambria Math font and subscripts for indices written with underscores to ensure cross-platform rendering (e.g. B_*i*(*t*), M_{*i*→*j*}(*t*), Bel_*j*(*t*), π_*ij*(*t*)); Greek letters appear as ε, λ, η_*t*, Θ.

Proof obligations. A theorem is considered established at the sketch level if the conclusion follows from the stated axioms plus the scope conditions above, with any model-specific assumptions (e.g. linear f, bounded η_*t*) declared. We avoid introducing hidden assumptions that change the force of an axiom.

Independence testing. To argue that axioms are not redundant, we exhibit (conceptually or by toy models) counter-models in which one axiom fails while the others hold. Example: a scripted exchange with fixed turn order (violates A6) while A1–A5 remain satisfied; or a channel with perfect encryption of R(M) (violates A2) while A1, A3–A5 hold.

Model classes. Where beneficial, we reference: (1) dynamic epistemic logic/action-model updates for A1–A2; (2) information-theoretic proxies (e.g. Bayesian surprise) as empirical indicators of belief change; (3) linear/switching state-space models for π_*ij*(*t*) under A5; (4) signaling-game intuitions for A3/A10 regarding code choice and leakage. For A4 and A7, we draw on conversation-analytic insights regarding turn design and repair organization (Heritage, 1984; Schegloff *et al*., 1977), operationalizing sequential structures as observable punctuation and repair moves.

Operationalization map. Each axiom/theorem is paired with candidate measures and manipulations (latency bands for A6; repair initiation rates for A7; grounding moves for A8; relevance deviations for A9; paralinguistic/physiological proxies for λ in A10). These proposals are not part of the proofs but are provided to enable falsification.

Statistical posture. When empirical tests are run, we anticipate entropy/variance comparisons for A3/A10, mixed-effects models for timing (A6) and repair (A7), and state-space estimation for π_*ij*(*t*) (A5). Uncertainty is reported with confidence/credible intervals; preregistered hypotheses are encouraged to avoid researchers' degrees of freedom.

Limitations. The framework deliberately abstracts from: (1) multi-party interactions beyond dyads, (2) higher-order beliefs ("I believe that you believe that I believe . . . "), (3) power structures embedded in institutional roles, and (4) cultural variation in pragmatic norms. These abstractions maintain tractability while preserving generalizability across domains (therapy, negotiation, organizational communication). Extensions addressing these factors must maintain level separation. Empirical validation will therefore depend on corpus-specific coding schemes, explicit construction of *R(M)*, observer-dependent punctuation assumptions *p_j*, and model-specific estimators for *p_j* (*r* |M) and π_*ij(t)*.

Results

Axioms (A1–A5): revised from Watzlawick et al

Here we recast the five classic "axioms" as genuine axioms: level-separated, weak enough to admit counter-models, and strong enough to yield non-trivial theorems. Rather than restating Watzlawick's propositions verbatim (Watzlawick *et al*., 1967), we specify minimal commitments about observability, interpretation, coding, segmentation, and a tractable relational state so that each claim has measurable antecedents and falsifiable consequents. The axioms below are therefore not descriptive slogans but constraints on models:

A1 fixes when behavior counts as communication (via belief update).

A2 makes relational content co-present with semantics.

A3 bounds interpretive variance by code choice.

A4 localizes attribution to agents' punctuation.

A5 links meta-signals to the evolution of dyadic stance.

Taken together, they form the smallest set we found that preserves the spirit of the original program while enabling proofs, metrics, and experimental interventions (cf. Bateson, 1972). Readers mainly interested in implications may skim the axiom statements and jump to the Theorems and the Applications section; each axiom is re-referenced there with concrete tests and examples.

A1. Unavoidability (in co-presence). If j can observe B_$i(t)$, then Bel_$j(t+1) \neq$ Bel_$j(t)$ Silence/inaction is still a behavior that can update beliefs.

A2. Meta-inevitability. For any *uptake M*, the receiver's interpretation includes a relational/meta dimension R(M) alongside content C(M). There is no *purely content-only uptake*: relational stance is inferred/attributed by participants from the action schema and context.

A3. Multi-code realizability. Any message can be realized in digital, analog, or mixed codes; code choice constrains interpretive variance across receivers.

A4. Subjective punctuation. Agents' segmentation functions p_i may differ; thus, the same behavior stream can be parsed into different causal/response episodes. Here "context" is not a pregiven environment but an operative punctuation by participants—an enacted narrative of "before/after" that grounds attributions of initiation, causality, and responsibility.

Why it matters: Divergent punctuation predicts rational reciprocal blame and escalation risk (see T2/T6).

Triferentiation Mapping: Responding to a reviewer's triad—position (role/asymmetry/legitimacy), function (sequential job: regulating/repairing/closing), and sense (symbolic horizon)—maps in our framework to: position ↔ relational state $\pi\_ij(t)$ and rights/obligations; function ↔ timing/turn-taking and repair (A6–A7); sense ↔ content frames C(M)and code/medium choices (A3, A10).

A5. Relational patterning (over repeated interaction). Statement. For dyad (i, j) engaged in repeated interaction ($T \geq T\_\min$, where $T\_\min$ is a minimum threshold for pattern formation, typically $T\_\min \geq 3$ exchanges), the relational state $\pi\_ij(t) \in \mathbb{R}$ encodes relative dominance/affiliation. At time t, the dyad is classified as:

(1) *Symmetric* if $|\pi\_ij(t)| \leq \varepsilon$ (near equality)

(2) *Complementary* if $|\pi\_ij(t)| > \varepsilon$ (hierarchical)

The relational component R(M) of each message updates π_ij over time according to:

$$\pi\_ij(t+1) = \pi\_ij(t) + \mathrm{f}\left(\mathrm{R}\left(\mathrm{M}\_\{i \to j\}(t)\right)\right) + \eta\_t$$

where f extracts dominance/leveling signals from R(M) and η_t is bounded noise.

Scope conditions:

(1) *Repeated Interaction Requirement:* A5 applies only when agents interact multiple times ($T \geq T\_\min$). For one-shot exchanges, π_ij is undefined or fixed at initial value $\pi\_ij(0)$ with no update dynamics.

(2) *Observable Meta-Signals:* Updates to π_ij require that R(M) contains extractable dominance/affiliation cues. If R(M) is systematically encrypted or suppressed (violating A2), π_ij remains constant.

(3) *Relational Memory:* Agents must maintain some representation of prior interactions for pattern stabilization. Amnesia or complete context loss resets π_ij to baseline.

Why these conditions matter: Without repeated interaction, terms like "stabilization" and "pattern" are meaningless, i.e. a single exchange cannot exhibit a pattern. The counter-model in Counter-Model 5 (Table 1) shows that one-shot communication satisfies A1-A4 but not A5, confirming that relational dynamics are a distinct commitment requiring temporal depth.

Table 1. Summary axiom independence: counter-models table

Axiom	Counter-model	Key feature	A1	A2	A3	A4	A5
A1 fails	Zero observability (encryption)	No belief update possible	✗	✓	✓	✓	✓
A2 fails	Pure task-bot	No relational layer	✓	✗	✓	✓	✓
A3 fails	Fixed-code constraint	Only one encoding	✓	✓	✗	✓	✓
A4 fails	Synchronized protocol	$p\_i \equiv p\_j$ enforced	✓	✓	✓	✗	✓
A5 fails	One-shot exchange	No relational history	✓	✓	✓	✓	✗

Operationalization under scope: Empirical tests of A5 should:

(1) Use dyads with $T \geq 3$–5 exchanges minimum

(2) Track $\pi\_ij(t)$ via continuous ratings or behavioral proxies (turn-share, interruption rates) across time

(3) Test whether manipulating R(M) (e.g. injecting dominance vs. leveling cues) predictably shifts π_ij trajectories

(4) Verify stabilization: after sufficient T, $|\pi\_ij(t)|$ should cross and remain above/below ε threshold

Levels. A1 (information), A2 (semantics/pragmatics), A3 (coding/medium), A4 (cognition/attribution), A5 (social dynamics). Designed to be consistent and as independent as possible.

Axiom independence: counter-models

Having established the five core axioms, we now demonstrate their logical independence. How to read this section: each construction shows exactly one axiom failing while the others hold, confirming non-redundancy. To verify that A1-A5 are not redundant, we exhibit counter-models where each axiom fails independently. Table 1 summarizes the results; below we sketch each construction.

Counter-model 1: A1 fails, A2-A5 hold. Scenario. Perfect encryption with zero observability.

Setup. Agents i and j interact via a channel where all behaviors $B\_i(t)$ are encrypted such that j observes only noise indistinguishable from baseline. Formally: j's observation function returns constant noise regardless of $B\_i(t)$.

Why A1 fails. Since j cannot distinguish $B\_i(t)$ from background, no belief update occurs: $Bel\_j(t+1) = Bel\_j(t)$. Communication criterion violated.

Why A2-A5 hold.

(1) *A2:* If a message *were* interpretable, it would still carry R(M). The encryption blocks interpretation, not the structure of messages.

(2) *A3:* Messages could still be encoded digitally/analogically; the encryption is downstream of encoding.

(3) *A4:* Each agent still has segmentation functions P_i, P_j (even if useless under encryption).

(4) *A5:* Relational state $\pi\_ij(t)$ could still be tracked (though it wouldn't update without observable meta-signals).

Conclusion. A1 is independent, i.e. it's the only axiom requiring actual observability for belief update.

Counter-model 2: A2 fails, A1, A3-A5 hold. Scenario. Pure task-bot communication with no relational layer.

Setup. Two AI systems exchange database queries in a formal language where R(M) is structurally undefined, i.e. only C(M) exists. Example: "SELECT * FROM *users* WHERE $id = 5$" has content but no stance, rights-duties, or affect encoding.

Why A2 fails. No relational component R(M) exists or can be extracted. Messages are purely propositional.

Why A1, A3-A5 hold.

(1) *A1:* Belief updates still occur (about database state).

(2) *A3:* Queries use digital encoding (symbolic language).

(3) *A4:* Systems could segment the query stream differently (e.g. treating nested queries as single vs. multiple episodes).

(4) *A5:* Not applicable to this pair (no dyadic relational state for bots), but if we insisted on tracking π_ij, it would remain at 0 (purely symmetric) since no meta-signals exist.

Conclusion. A2 is independent, i.e. relational content is not logically necessary for information transfer.

Counter-model 3: A3 fails, A1-A2, A4-A5 hold. Scenario. Fixed-code constraint—only one encoding available.

Setup. Agents communicate via a channel that supports *only* digital encoding (e.g. text-only chat with no formatting, no timing information, no paralinguistic cues). Alternatively: only analog (e.g. pure pantomime with no discrete symbols).

Why A3 fails. Messages cannot be realized in multiple codes, i.e. the claim "any message can be realized in digital, analog, or mixed codes" is false. Only one realization exists per message.

Why A1-A2, A4-A5 hold.

(1) *A1:* Belief updates still occur from text messages.

(2) *A2:* Text still carries R(M) (e.g. "Please help" vs. "Help me" vs. "HELP" convey different stances even in pure text).

(3) *A4:* Agents can segment text streams differently.

(4) *A5:* Meta-signals in text (e.g. formality, brevity) still update $\pi\_ij(t)$.

Conclusion. A3 is independent, i.e. it's possible to have communication without multi-code realizability.

Counter-model 4: A4 fails, A1-A3, A5 hold. Scenario. Synchronized punctuation via shared protocol.

Setup. Agents interact under a rigid protocol that enforces identical segmentation. Example: formal debate with explicit turn labels ("Opening statement," "Rebuttal 1," etc.) where both agents must use the same episode boundaries. Formally: $p\_i \equiv p\_j$ by design.

Why A4 fails. The axiom claims agents' segmentation functions *may differ*. In this counter-model, they cannot differ, i.e. $p\_i = P\_j$ is enforced, so the possibility of divergence is eliminated.

Why A1-A3, A5 hold.

(1) *A1:* Belief updates occur from observing turns.

(2) *A2:* Each turn carries relational content (e.g. deference, challenge).

(3) *A3:* Turns can be encoded digitally (written) or with mixed codes (spoken with prosody).

(4) *A5:* Meta-signals accumulate (e.g. repeated interruptions in one direction → complementarity).

Conclusion. A4 is independent, i.e. identical punctuation is consistent with all other axioms.

Counter-model 5: A5 fails, A1-A4 hold. Scenario. One-shot exchange ($T = 1$)

Setup. Two strangers exchange a single message and never interact again. For example: a tourist asks a passerby for directions, receives an answer, and they part ways. Since $T = 1 < T\_\min$, A5's scope conditions are not satisfied—there is no repeated interaction for relational pattern formation.

Why A5 fails. The axiom explicitly requires $T \geq T\_\min$ (repeated interaction) for relational dynamics to apply. With only one exchange, the relational state π_ij has no trajectory to evolve along and cannot stabilize into symmetric or complementary patterns. The update equation $\pi\_ij(t+1) = \pi\_ij(t) + f(R(M)) + \eta\_t$ is never iterated, so terms like "stabilization" and "pattern" are undefined. A5's central claim, i.e. that repeated meta-signals push $|\pi\_ij|$ across threshold ε cannot be tested with a single data point.

Why A1-A4 hold.

(1) *A1* (*Unavoidability*): The single message still updates beliefs. The tourist's $Bel\_j(t)$ changes from uncertainty about directions to knowledge of the route.

(2) *A2* (*Meta-inevitability*): The message carries relational content R(M). Even "Turn left at the corner" conveys stance (e.g. helpfulness, formality level, patience/impatience via tone).

(3) *A3* (*Multi-code realizability*): The direction can be given digitally (words), analogically (pointing gesture), or mixed (words + gesture + prosody).

(4) *A4* (*Subjective punctuation*): Each agent could segment the encounter differently. The tourist might view it as "I initiated by asking," while the passerby might view it as "They interrupted my walk, I responded."

Conclusion. A5 is independent because it requires temporal depth (repeated interaction) that A1-A4 do not mandate. One can have communication (A1), with relational content (A2), using various codes (A3), segmented differently by participants (A4), without any relational pattern formation (A5), i.e. simply by restricting to $T = 1$.

These counter-models establish that no axiom is derivable from the others. Each makes a distinct commitment: A1 (observability for update), A2 (relational inevitability), A3 (encoding flexibility), A4 (punctuation variability), A5 (relational dynamics over time). Together they span the minimal constraints needed to capture Watzlawick's original insights.

Additional axioms (A6–A10): timing, repair, common ground, cooperation, channels

The original five axioms describe inevitability, meta-content, coding, punctuation, and relational patterning. Many interaction failures, however, hinge on temporal coordination, repair mechanisms, shared context, default cooperativity, and the gap between public signals and private states. We therefore propose the following additions. Each axiom is framed to be (1) level-compatible with A1–A5, (2) operationalizable for measurement, and (3) weak enough to admit counter-models (so independence tests are possible).

These organizational axioms (timing, repair, common ground, relevance default, public–private channels) connect directly to coding schemes and experimental manipulations used in conversation analysis and human–AI interaction, as detailed in the empirical agenda.

A6. Turn-taking and timing (temporal coordination). Statement. There exist timing norms Θ(latency bands, overlap tolerances) such that, for observable turns $B\_i(t)$, deviations from Θ increase the posterior variance of $I\_j(\cdot)$ over C(M) and R(M), and alter floor-rights attribution.

Intuition. Timing is not just packaging; it shapes *who has the floor* and thus the relational interpretation.

Operationalization. Measure response latencies and overlaps; model their effect on (a) comprehension error rates and (b) perceived dominance/affiliation.

Independence Note. A6 can fail (e.g. fully scripted exchanges) while A1–A5 hold.

A7. Repairability (error-handling). Statement. If agent *j*'s uncertainty about C(M) or R*(M)* exceeds a threshold τ after observing B_*i*(*t*), then (in the absence of inhibiting costs) there exists an admissible *repair move* ρ such that B_*j*(t+1) = ρ(M{$i \to j$}(*t*)) reduces that uncertainty.

Intuition. Systems of interaction include built-in mechanisms to fix misunderstandings (clarifications, repeats, reformulations).

Operationalization. Annotate repair initiations/resolutions; estimate uncertainty reduction pre/post repair.

A8. Common ground (context maintenance). Statement. Let CG(*t*) denote mutually presupposed propositions at time *t*. Successful uptake requires either (1) C(M) $\subseteq$ CG(*t*) or (2) an update move μ that revises CG to CG(t+1) $\supseteq$ C(M) without contradiction.

Intuition. Content uptake is constrained by what is already mutually believed; when content outruns CG(*t*), grounding moves are needed.

This axiom builds on Clark and Brennan's (1991) grounding model, which specifies how conversational partners establish mutual understanding through grounding acts (acknowledgments, repairs, expansions). Their framework has been extended to human-AI interaction (Brennan, 1998; Skantze, 2021) to address the challenge that artificial agents often lack human-like common ground accumulation—leading to failures our A8 predicts: utterances interpretable in isolation but unintegrable into dialogue context. Stalnaker's (2002) presupposition logic provides formal machinery for modeling CG(*t*) as a context set (possible worlds compatible with mutual assumptions); our A8 can be instantiated using Stalnaker's update rules, where grounding moves μ are context-change potentials that eliminate incompatible worlds. This formalization is especially valuable for socio-technical systems (human-AI, multi-agent), where CG(*t*) may be asymmetric or incomplete and grounding failures trigger repair cascades (A7).

Operationalization. Code presupposition triggers/grounding moves; track acceptance vs. challenge.

A9. Relevance/cooperation (defeasible default). Statement. By default, agents choose contributions that maximize expected relevance given current goals and CG(*t*); violations are licensed by context and are themselves informative about R(M).

Intuition. Gricean cooperativity is not an axiom of truthfulness but a *prior* about design.

Contemporary formal accounts refine this cooperativity principle: relevance theory (Sperber and Wilson, 1986/1995) formalizes how communicators optimize for cognitive effects relative to processing effort, providing a decision-theoretic basis for the relevance default we posit in A9. Game-theoretic pragmatics (Franke and Jäger, 2016) extends this further by modeling speakers and hearers as Bayesian rational agents playing signaling games, where interpretive variance (central to our A3) emerges from equilibrium selection under different cost structures and common priors. These frameworks provide natural operationalizations for our "expected relevance" and show how A9's defeasibility follows from context-dependent utility functions rather than categorical rules.

Operationalization. Manipulate task goals; test whether off-topic or over-informative messages shift R(M) (e.g. perceived dominance, withdrawal).

A10. Public vs private channels (signaling vs state). Statement. Each agent has a private state S_*i*(*t*) (beliefs, intentions, affect). Public signals M are generated from S_*i*(*t*) through an encoding policy κ_*i*. Leakage λ (e.g. prosody, micro-expressions) causes partial, sometimes unintended, mapping of S_*i* into R(M).

Intuition. Misalignment between what is signaled and what is privately held is common—and diagnostic.

Operationalization. Combine content analysis with paralinguistic/physiological measures to estimate λ; test attribution errors when κ_*i* is strategic.

Short consequences and test points.

(1) *From A6: Latency dominance effect.* Short, frequent entries into the floor should increase $|\pi\_ij|$ (toward complementarity).

(2) *From A7: Repair deficit pathology.* Suppressing repair (raising costs) should raise misattribution rates and escalate blame chains.

(3) *From A8: Grounding bottleneck.* Introducing new terms without grounding should raise failure-to-uptake probabilities, even with perfect coding.

(4) *From A9: Relevance as meta-signal.* Irrelevance raises the probability that R(M) is interpreted as disengagement or dominance (context-dependent).

(5) *From A10: Leakage asymmetry.* Increased leakage λ reduces plausible deniability; strategic encodings with low λ increase interpretive variance across receivers.

Independence test kit for A6–A10

A6. Turn-taking and timing. Counter-model (A6 false; A1–A5, A7–A10 hold). Scripted, latency-free chat where turns are queued and delivered at fixed ticks regardless of response time (no floor competition; overlaps impossible). Belief updates, meta-content, punctuation, and π dynamics still operate; repair (A7) is available, common ground (A8) updates, etc.

Positive model (A6 true adds unique predictions). Same content and codes, but manipulate inter-turn latencies/overlap permissions. Prediction unique to A6: latency/overlap shifts attributions of floor rights and perceived dominance, even when C(M), R(M), and p_i are held constant.

Falsification Signature. If varying latency bands doesn't change floor attribution or dominance judgments (given power to detect), A6 is redundant.

A7. Repairability. Counter-model (A7 false; others hold). Channel prohibits clarification (no backchannels; repair moves rejected by design). Everything else (A1–A6, A8–A10) intact.

Positive model (A7 true adds unique predictions). Introduce low-cost repair operators ρ; misinterpretations drop conditional on repair with no change to code or timing.

Falsification Signature. If enabling repair moves doesn't reduce uncertainty/misunderstanding (holding A3/A6 constant), A7 is not independently needed.

A8. Common ground. Counter-model (A8 false; others hold). Zero shared lexicon/frames; all content is novel and cannot be integrated into a shared set CG(t). Agents still update beliefs (A1), infer relational stance (A2), use codes (A3), segment (A4), and π_{ij} evolves (A5), but content uptake routinely fails.

Positive model (A8 true adds unique predictions). Prime a shared glossary or seed CG(t); with identical messages, uptake rises and need for repair (A7) falls—*even if* timing (A6) and code (A3) remain unchanged.

Falsification Signature. If establishing CG(t) doesn't improve uptake beyond what A3/A7 can explain, A8 is not adding independent bite.

A9. Relevance/cooperation. Counter-model (A9 false; others hold). Agents optimize an adversarial or random objective (produce off-goal turns by default). A1–A8, A10 still function; you will observe communication and relations, but baseline contribution design is non-cooperative.

Positive model (A9 true adds unique predictions). With the same task and CG(t), cooperative priors yield systematically shorter paths to task goals and different default interpretations of ambiguous acts (e.g. off-topic contributions read as affiliative rather than hostile).

Falsification Signature. If flipping the contribution prior (cooperative $\leftrightarrow$ non-cooperative) doesn't change relevance ratings, uptake, or inferred R(M) (controlling A3/A8), A9 is dispensable.

A10. Public vs private channels (leakage) λ. Counter-model (A10 false; others hold). Perfectly sealed encoding: public M is a deterministic function of content; relational state and private affect never leak (fix $\lambda = 0$). A1–A9 remain true; you still get belief change, coding, punctuation, π dynamics, but no differential effect of paralinguistic/side channels.

Positive model (A10 true adds unique predictions). Vary λ by constraining channels (text-only ↔ audio-video ↔ audio-video + physiological). Predict the inverted-U in posterior interpretive entropy $\overline{H}$ vs. λ, and leakage-driven shifts in perceived sincerity/dominance at fixed C(M) and code (A3).

Falsification signature: If manipulating leakage leaves $\overline{H}$ and R(M) unchanged once A3 is controlled, A10 is redundant.

Core theorems (with short proofs)

In this theoretical paper, the named theorems (T1–T7) constitute our primary results. Now we are ready to demonstrate that the axioms are not merely redescriptions but generators of testable consequences. The following theorems translate everyday phenomena, such as silence, disagreements about "who started it," and stable dominance/equality patterns, into entailments of A1–A5, accompanied by explicit proof sketches. Each proof is intentionally minimal: enough to show that the conclusion follows from the axioms plus mild regularity assumptions, and precise enough to suggest operational measures and interventions. Together, they illustrate our program: move from compact axioms to falsifiable predictions and, ultimately, to experimental designs that can adjudicate between competing models of interaction. Readers can skip the algebra and still follow the practical reading indicated after each theorem.

T1. Communicativity of silence. Claim. In co-presence, intentional silence/withdrawal by *i* communicates to *j*.

Proof sketch. Let $B_i(t) = \varnothing$ (silence/withdrawal) be observable to j in context where a response was expected. By the belief representation (§Methods), agent *j* holds prior beliefs $Bel_j(t)$ including propositions about i's cooperativeness, engagement, and stance.

In context (e.g. *j* asked a direct question), the absence of expected behavior carries information. Specifically, *j*'s update operator U_j must revise confidence scores:

(1) Propositions like ("*i* is cooperative", *c*) → ("*i* is cooperative", c – δ) where $\delta > \varepsilon$

(2) New propositions may enter: ("*i* is withholding", c') with $c' > \theta$

(3) Or: ("topic is face-threatening", c'') with $c'' > \theta$

Since $|\Delta c| > \varepsilon$ for at least one proposition, we have $Bel_j(t+1) \neq Bel_j(t)$. By A1's communication criterion, agent *i* has communicated. The content of the communication (refusal, disengagement, disapproval) depends on context, but the fact of communication follows necessarily from observability + belief update. ∎

T2. Punctuation asymmetry ⇒ reciprocal blame. Claim. If segmentation functions $p_i \neq p_j$, there exist behavior streams where each agent rationally attributes escalation to the other.

Proof sketch. Construct a three-step sequence:

(1) $t = 1$: $B_i(1)$ = *micro-snub* (e.g. curt response)

(2) $t = 2$: B_j(2) = brusque reply

(3) $t = 3$: B_i(3) = defensive statement

Agent *i*'s segmentation: p_i cuts the stream at $t = 2$, treating $B_j(2)$ as initiation. Thus i's beliefs update:

(1) Bel_i(before $t = 2$) = {("*j* is cooperative", 0.7)}

(2) Observe B_*j*(*2*) → Bel_*i*(after $t = 2$) = {("*j* initiated hostility", 0.8), ("*j* is uncooperative", 0.7)}

(3) B_*i*(3) is framed as defensive response

Agent *j*'s segmentation: *p_j* cuts at $t = 1$, treating B_*i*(1) as initiation. Thus:

(1) Bel_*j*(before $t = 1$) = {("*i* is cooperative", 0.7)}

(2) Observe B_i(1) → Bel_j(after $t = 1$) = {("*i* initiated hostility", 0.8), ("*i* is uncooperative", 0.7)}

(3) B_*j*(2) is framed as defensive response

From A4, each agent's attribution follows from their segmentation function applied to the same objective stream. Since content C(M) is constant but segmentation differs, the divergent blame attributions are rational consequences of $p\_i \neq p\_j$, not evidence errors. ∎

T3. Meta-signals lock in symmetry or complementarity. Claim. Repeated dominance/submission meta-signals push $|\pi\_ij|$ above/below ε, stabilizing complementary/symmetric patterns.

Proof sketch. Let the relational state evolve by:

$$\pi\_ij(t+1) = \pi\_ij(t) + \mathrm{f}\big(\mathrm{R}\big(M\_\{i \to j\}(t)\big)\big) + \eta\_t$$

where f extracts the dominance/leveling signal from the relational component R(M), and η_t is bounded noise.

For each message M_{*i*→*j*}(*t*), agent j updates beliefs about the relationship:

(1) Dominance displays → Bel_*j* includes ("*i* is asserting control", *c*) with high $c \rightarrow$ f(R(M)) > 0

(2) Leveling/affiliation cues → Bel_*j* includes ("i seeks equality", *c*) → f(R(M)) < 0

Over repeated interactions (*T* iterations), if dominance displays predominate:

(1) E[f(R(M))] > δ for some δ > 0

(2) By linearity and bounded noise: $\mathrm{E}[\pi\_ij(T)] \approx \pi\_ij(0) + T \cdot \delta$

(3) For sufficiently large *T*: $|\pi\_ij(T)| > \varepsilon$ (crosses complementarity threshold)

Conversely, if leveling predominates (E[f(R(M))] < 0), $\pi\_ij(T) \rightarrow 0$, stabilizing symmetry ($|\pi\_ij| \leq \varepsilon$).

From A5, the relational pattern (symmetric vs. complementary) is determined by accumulated meta-signals R(M) over time, which shape participants' beliefs about relative standing and are reflected in the state variable π_ij.

Optional formalization. Use dynamic epistemic logic for belief updates (A1–A2) and a simple linear state-space model for π_ij dynamics (A5). ∎

T4. Early-anchor vulnerability (A6+A8). Claim. If an early message set CG(*t_0*) with an incorrect anchor and turn-taking latencies keep later corrections outside Θ, then posterior beliefs will disproportionately weight the anchor, increasing the probability of persistent misunderstanding.

Informal Sketch. Latency constraints limit effective repair windows; anchoring biases propagate through *CG*.

T5. Credibility amplification via code choice (A3+A10). Claim. For the same C(M), analog-rich encodings with moderate leakage λ yield lower interpretive entropy about R(M) than purely digital encodings, up to a leakage ceiling beyond which noise dominates.

Proof Sketch. Let H_bar(M) denote mean posterior interpretive entropy across receivers (defined in Methods). We establish a partial order:

Step 1: Digital vs. Analog-rich (A3). By A3, code choice constrains interpretive variance. Digital codes use discrete symbols with high arbitrariness (word choice, punctuation) but low redundancy. Analog codes embed graded cues (prosody, gesture) with lower arbitrariness but higher redundancy across channels. For R(M) extraction: analog codes provide multiple correlated signals (pitch + volume + facial expression all encoding affect), while digital codes provide single symbolic channel. By information theory, correlated redundant signals reduce posterior uncertainty when aggregated:

$$\text{H\_bar}(M \mid \text{mixed}) \leq \text{H\_bar}\,(M \mid \text{digital} - \text{only})$$

where equality holds only if analog channels are completely uninformative.

Step 2: Leakage effect (A10). By A10, leakage λ quantifies unintended mapping of private state S_*i* into R(M). Define three regimes:

- Low leakage ($\lambda \approx 0$): Strategic encoding succeeds; R(M) is sparse but intentional. High interpretive variance because receivers must infer from minimal cues: H_bar(M $|\lambda \sim 0$) is large.
- Moderate leakage ($\lambda \approx \lambda^*$): Partial leakage provides diagnostic cues without overwhelming noise. Receivers integrate intentional signals with leaked affect/stance markers, reducing uncertainty:

H_bar(M $|\lambda \approx \lambda^*$) is minimized.

(1) High leakage ($\lambda >> \lambda^*$): Uncontrolled leakage introduces noise (conflicting signals, random physiological artifacts). Receivers cannot distinguish signal from noise: H_bar (M $|\lambda >> \lambda^*$) increases again. Formally, H_bar(M $|\lambda$) follows inverted-U:

∂H_bar/$\partial\lambda < 0$ for $\lambda < \lambda^*$ (increasing leakage reduces entropy)

∂HH_bar/$\partial\lambda > 0$ for $\lambda > \lambda^*$ (excessive leakage increases entropy)

with minimum at $\lambda = \lambda^*$, where λ^* depends on channel reliability and receiver inference capacity.

This ordering concerns posterior dispersion under different code/leakage regimes; it is descriptive rather than normative and does not evaluate relational quality.

Step 3: Combined ordering. Combining A3 and A10:

$$\begin{aligned}\text{H\_bar}(M \mid \text{mixed}, \lambda \approx \lambda^*) &< \text{H-(M} \mid \text{analog} - \text{only}, \lambda \approx \lambda^*) \\ &< \text{H\_bar}(M \mid \text{digital} - \text{only}, \lambda \approx 0)\end{aligned}$$

The inequality chain holds because:

(1) Mixed codes provide redundancy (A3) plus diagnostic leakage (A10)

(2) Analog-only lacks discrete symbolic grounding

(3) Digital-only suppresses paralinguistic cues that reduce ambiguity

Note: The "minimum" in the leakage curve is a descriptive feature of uncertainty under the model, not a normative objective; lower entropy does not entail better interactional outcomes.

Prediction. Experimentally vary code (digital/analog/mixed) and channel richness (text/audio/video). Measure H_bar via between-rater variance on R(M) coding. Predict ordering above, with minimum entropy at mixed codes with moderate channel richness. ∎

Empirical note. The leakage optimum λ* is context-dependent. In high-stakes deception (poker, negotiation), low λ is strategic; in trust-building (therapy, team formation), moderate λ signals authenticity and reduces interpretive entropy by providing corroborating cues.

T6. Symmetrical escalation: when axioms generate pathology. The theorems above (T1-T3) demonstrate that our axioms generate recognizable interaction patterns, e.g. silence communicates, punctuation differences yield blame, meta-signals stabilize relationships. But Watzlawick's deepest contribution was identifying *pathological* patterns: self-perpetuating cycles that trap participants despite their best intentions. We now show that the same axioms, under specific scope conditions, generate symmetrical escalation (Bateson's "schismogenesis"), i.e. a feedback loop where rational agents, each correctly following their own punctuation, drive the dyad toward increasing conflict. This theorem illustrates how formalization moves beyond description to diagnosis: by identifying the axioms responsible (A4, A5, A7), we pinpoint where interventions must target to break the cycle.

Statement. Punctuation-Blame Spiral (Symmetrical Escalation)

Claim. When agents with divergent punctuation ($p\_i \neq p\_j$) engage in repeated interaction under conditions of (1) high reactivity to perceived aggression, (2) no repair mechanisms (A7 suppressed), and (3) symmetric relational positioning ($|\pi\_ij(0)| \leq \varepsilon$), the system exhibits symmetrical escalation: mutual attributions of blame intensify over time, and either (1) $|\pi\_ij(t)|$ grows without bound, or (2) behavioral intensity (measured by dominance displays in R(M)) oscillates with increasing amplitude until external intervention or relationship termination.

Proof Sketch.

Setup. Let agents i and j interact over T rounds with:

Divergent punctuation: $p\_i \neq p\_j$, specifically with offset segmentation (i's episode k begins at j's episode k-1 midpoint)

Initial symmetry: $\pi\_ij(0) = 0$ (near-equals at start)

Reactivity assumption: Each agent responds to perceived aggression with proportional counter-aggression. Formally: if agent j infers from segmentation that i initiated at t, then R(M_{$j \to i$}(t+1)) contains dominance signal of magnitude $\alpha \cdot$ f(R(M_{$i \to j$}(t))) where $\alpha > 1$ (amplification factor)

No repair: is disabled, i.e. agents cannot clarify punctuation or check attributions.

Step 1: Reciprocal blame (*from T2*). By T2, divergent punctuation ensures each agent rationally attributes initiation/escalation to the other. At $t = 1$:

Agent i observes B_j(1) and, segmenting via p_i, believes: ("j initiated hostility", $c > \theta$)

Agent j observes B_i(1) and, segmenting via p_j, believes: ("i initiated hostility", $c > \theta$)

Step 2: Reactive counter-aggression. Given blame attributions, each responds at $t = 2$ with R(M) containing dominance/challenge signals:

i encodes f(R(M_{$i \to j$}(2))) > 0 (dominance display, perceived as "defending against j's aggression")

j encodes f(R(M_{$j \to i$}(2))) > 0 (symmetric response)

Step 3: Feedback amplification. At $t = 3$, each observes the other's counter-aggression and, *still operating under divergent punctuation*, interprets it as escalation rather than defense:

Bel_i includes ("j is escalating", c') with $c' > c$.

Bel_j includes ("i is escalating", c') with $c' > c$.

Given reactivity $\alpha > 1$:

$$\mathrm{f}\big(\mathrm{R}\big(M\_\{i \to j\}(3)\big)\big) = \alpha \cdot \mathrm{f}\big(\mathrm{R}\big(M\_\{j \to i\}(2)\big)\big) \;>\; \mathrm{f}\big(\mathrm{R}\big(M\_\{j \to i\}(2)\big)\big)$$

$$\mathrm{f}\big(\mathrm{R}\big(M\_\{j \to i\}(3)\big)\big) = \alpha \cdot \mathrm{f}\big(\mathrm{R}\big(M\_\{i \to j\}(2)\big)\big) \;>\; \mathrm{f}\big(\mathrm{R}\big(M\_\{i \to j\}(2)\big)\big)$$

Step 4: Relational state trajectory. From A5, π_ij evolves as:

$$\pi\_ij(t+1) = \pi\_ij(t) + \mathrm{f}\big(\mathrm{R}\big(M\_\{i \to j\}(\mathrm{t})\big)\big) - \mathrm{f}\big(\mathrm{R}\big(M\_\{j \to i\}(\mathrm{t})\big)\big) + \eta\_t$$

In symmetrical escalation, the dominance signals are roughly balanced: f(R(M_{$i \to j$})) ≈ f(R(M_{$j \to i$})) at each step, so the *net* π_ij oscillates near 0 (remains symmetric). However, the *magnitude* of each signal grows: |f(R(M_{$i \to j$}(t)))| increases over iterations.

Outcome (a)*: Unbounded intensity.* If $\alpha > 1$ persists and η_t is bounded, then:

$$\mathrm{E}[|\mathrm{f}(\mathrm{R}(M(t)))|] \approx |\mathrm{f}(\mathrm{R}(M(0)))| \cdot \alpha\text{^}t \to \infty \text{as } t \to \infty$$

(Note: this holds when signals are independent or weakly dependent). Behavioral intensity (e.g. volume, interruptions, insults) escalates indefinitely until physical/social constraints halt interaction.

Outcome (b)*: Oscillating amplitude.* If there are soft bounds (e.g. social norms prevent overt aggression), the system oscillates between:

(1) Overt conflict (high |f(R(M))|)

(2) Brief détente (temporary reduction)

(3) Re-escalation triggered by minor provocation (punctuation divergence still active)

Amplitude increases: each conflict peak is more intense than the last.

Why self-correction fails. Without repair (A7), agents cannot:

(1) Recognize punctuation divergence ("Wait, do you think *I* started this?")

(2) Revise segmentation to align P_*i* and P_*j*

(3) Update blame attributions ("Maybe we both contributed")

The system is *locally rational* (each agent is responding appropriately to their perceived reality) but *globally pathological* (the dyad is stuck in escalation).

Intervention points (from axioms):

(1) *Synchronize punctuation* (*A4*): External mediator imposes shared timeline ("Let's agree: the conflict started when *X* happened")

(2) *Enable repair* (*A7*): Lower cost of clarification ("Can you explain what you meant?")

(3) *Interrupt feedback* (*A5*): Introduce cooling-off period (break the iteration)

(4) *Reframe meta-signals* (*A2*): Teach leveling communication to reduce f(R(M))

Worked example: Inter-communal border dispute and Structured Democratic Dialogue intervention. Two neighboring communities (A and B) share a contested border zone historically used by both for grazing and water access. Recent droughts have increased resource pressure, while minimal governmental mediation leaves communities to manage tensions bilaterally. Initial mutual respect deteriorates through punctuation divergence that neither community recognizes.

Setup. Bel_A(0) = {("Community B respects boundaries", 0.6), ("conflict can be avoided", 0.7)}; Bel_B(0) = {("Community A respects boundaries", 0.6), ("conflict can be avoided", 0.7)}.

Segmentation functions diverge: *p*_A segments the conflict from "B's boundary crossing," *p*_B segments from "A's aggressive response." Reactivity $\alpha = 1.4$ (scarcity amplifies threat perception). Repair mechanisms unavailable: no neutral mediator, direct communication viewed as weakness.

T = 1: Community B's herders move 50 cattle across informal boundary to reach water during drought. Community A interprets via P_A: ("B is initiating land grab", 0.8), ("B is testing our resolve", 0.75) → f(R(M_{B→A}(1))) = +0.5 (territorial challenge detected).

T = 2: A sends 15 youth to boundary zone; they shout warnings and throw stones near B's herders (no injuries). Community B, who coded cattle movement as traditional resource access, interprets via P_B: ("A initiated unprovoked attack", 0.8), ("our community is under threat", 0.75) → f(R(M_{A→B}(2))) = +0.7 (violence detected).

T = 3: B mobilizes 25 youth to "defend rights"; they establish armed camp in contested zone. A, still operating under P_A (cattle crossing = initiation), sees this as major escalation: ("B is preparing to attack", 0.75), ("we must defend our land", 0.85) → f(R(M_{B→A}(3))) = +1.2.

T = 4–7: Intensity grows exponentially through reactivity $\alpha = 1.4$: $|f(R(M))| = 0.5 \rightarrow 0.7 \rightarrow 0.98 \rightarrow 1.37 \rightarrow 1.92 \rightarrow 2.69 \rightarrow 3.77$. Road blockade ($T = 4$), property damage ($T = 5$), physical altercations ($T = 6$), injuries requiring medical attention ($T = 7$). National media attention begins. Relational state π_AB remains near 0 (both communities feel victimized; neither achieves dominance). Mutual blame: "You started this."

Why self-correction failed: No repair channel available (A7 suppressed); $p\_A \neq p\_B$ remains unrecognized by participants; each community rational given its segmentation; scarcity conditions sustain $\alpha > 1$; absence of governance infrastructure removes intervention capacity.

Intervention via Structured Democratic Dialogue (SDD): Formalizing A4 Synchronization. The Structured Democratic Dialogue process, developed and applied extensively in Cyprus to bridge Greek-speaking and Turkish-speaking communities (Broome, 1998, 2002; Broome and Anastasiou, 2011, AUTHORS), provides a systematic operationalization of punctuation repair (A4). The SDD methodology maps directly onto our axiom framework:

(1) *Phase 1: Triggering question.* Neutral facilitators pose: "What are the key events in this conflict?" Each community generates its event list independently. This phase makes latent segmentation functions P_A and P_B observable: Community A lists "boundary crossing by B" as Event 1; Community B lists "stone-throwing by A" as Event 1. The divergence is implicit but documented.

(2) *Phase 2: Clarification.* Participants explain each event's meaning in their own terms. This exposes the interpretation functions I_A and I_B: A explains why boundary crossing signals land-grab intent; B explains why it was routine drought-driven access. Critically, this occurs in structured format that prevents immediate counter-argument, lowering reactivity (reducing α).

(3) *Phase 3: Structuring via interpretive structural modeling.* Events are organized into causal/temporal patterns through iterative pairwise comparison. Participants must agree on precedence relations even when they disagree on causation. This forces alignment: the shared timeline reveals that P_A cuts at $T = 1$, P_B cuts at $T = 2$. The punctuation asymmetry becomes visible to all participants. Key update: Both communities revise beliefs to include ("We were responding to different provocations", 0.85), ("misunderstanding is a core driver", 0.8).

(4) *Phase 4: Interpretive session.* Root causes identified collaboratively. Participants recognize: "Our conflict escalated not because either side was irrational, but because we sliced the event stream differently." This meta-cognitive insight—understanding that $P\_A \neq P\_B$—breaks the attribution of malicious intent. Bel_An updates to reduce ("B wants to harm us", 0.9 → 0.3); Bel_B updates symmetrically.

Outcome. De-escalation begins within 48 hours of SDD session. Communities establish joint water-access protocol (addresses scarcity that elevated α). Elder-to-elder communication channel created (enables A7 repair for future events). Follow-up SDD sessions at $T+30$ days confirms sustained reduction in $|f(R(M))|$ and stabilization of π_AB near 0 with cooperative rather than conflictual symmetry.

SDD as formalized A4 intervention. The Cyprus applications (*vide supra*) demonstrate that SDD systematically addresses the $p\_i \neq p\_j$ problem driving symmetrical escalation. By externalizing segmentation functions (Phase 1–2), aligning them through structured comparison (Phase 3), and enabling meta-cognitive recognition of the divergence (Phase 4), SDD operationalizes exactly the punctuation repair our framework identifies as necessary. The method's effectiveness in bridging deeply divided communities, such as Greek and Turkish Cypriots navigating decades of conflict, provides empirical validation of T6's intervention logic. Where punctuation asymmetry remains hidden, escalation persists; where it is made visible and aligned, de-escalation becomes possible even in high-stakes inter-communal contexts.

Formal models of negotiation and conflict resolution provide complementary frameworks for understanding SDD's mechanism. Raiffa's (1982) negotiation analysis emphasizes creating shared reference points and synchronized problem representations—precisely what SDD's Phase 3 structuring accomplishes by aligning p_A and p_B. Axelrod's (1984) iterated cooperation models show how establishing reciprocity norms and reducing reactivity α stabilizes cooperation; SDD's multi-phase design operationalizes this by making meta-communication explicit (lowering reactivity through structured format) while building common ground incrementally (enabling reciprocity). Our framework formalizes the punctuation-repair mechanism (A4) that these earlier models implicitly relied upon: without synchronized event timelines, even rational cooperators cannot escape symmetrical escalation (T6).

Clinical/applied note. Symmetrical escalation is ubiquitous in:

(1) Marital conflict ("nagging-withdrawal" cycles where each sees themselves as responding to the other)

(2) International relations (arms races, tit-for-tat sanctions)

(3) Organizational disputes (department rivalries)

(4) Online flame wars

The formalization reveals that *neither party is irrational*, i.e. the pathology is systemic, arising from axiom interactions. This shifts intervention from "fix the bad actor" to "repair the punctuation/repair mechanisms."

Empirical predictions from T6

Prediction 1 (*Escalation signature*): Dyads with experimentally induced $p\_i \neq p\_j$ should show:

(1) Increasing interruption rates over trials

(2) Rising vocal intensity/pitch

(3) More blame language ("You started it," "You always . . . ")

(4) Mutual perception of being victimized

Prediction 2 (*Symmetry maintenance*)*:* Despite escalating intensity, π_ij should remain near 0 (symmetric) rather than crossing ε into complementarity, i.e. distinguishing symmetrical escalation from dominance struggles.

Prediction 3 (*Repair intervention*): Allowing one repair move per 3 exchanges should reduce escalation rate by ∼50% (testable in lab).

Prediction 4 (*Punctuation alignment*): After synchronizing p_i and p_j via timeline reconstruction task, blame attributions should converge and intensity should decrease within 2–3 subsequent exchanges.

T7. Meta-communication paradox: when denying dominance asserts it. The theorems above (T1-T6) demonstrate that our axioms generate both normative interaction patterns and pathological escalations. We now show that the axioms also capture a deeper class of pragmatic paradoxes: situations where explicit meta-communicative attempts to define the relationship produce effects opposite to their stated intent. Specifically, when an agent near a complementarity threshold attempts to verbally assert symmetry, the meta-communicative act itself signals dominance, thereby increasing rather than decreasing relational asymmetry. This theorem reveals how A2 (meta-inevitability) and A5 (relational dynamics) together generate self-referential traps characteristic of the double binds Bateson and Watzlawick identified as central to pathological communication.

Statement. Meta-Communicative Dominance Assertion.

Claim. For dyads in relational configurations near the complementarity threshold ($|\pi\_ij(t)| \approx \varepsilon$), explicit meta-communicative messages attempting to assert symmetry, i.e. statements with content C(M) = "I value equality" or "I'm not trying to control you," carry relational components R(M) that signal dominance, thereby pushing $\pi\_ij(t+1)$ further from symmetry. The effect intensifies with repetition: agents who repeatedly meta-communicate symmetry intentions generate stronger complementarity than those who remain silent about the relationship.

Proof Sketch.

Setup. Let dyad (i,j) have relational state $\pi\_ij(t) \approx \varepsilon$ (hovering near complementarity threshold). Agent i intends to reduce perceived dominance and believes that explicitly stating egalitarian intent will achieve this.

Agent i sends message M with:

(1) C(M) = "I want us to be equals" (or equivalent symmetry assertion)

(2) Intended f(R(M)) < 0 (leveling signal)

Step 1: Meta-communicative framing as dominance signal (from A2). By A2, every message carries relational content R(M) co-conveyed with C(M). The act of explicitly defining the relationship—unilaterally stating what kind of relationship "we" should have—is itself a relational move that asserts authority over the relational frame.

Specifically: The speech act of relationship-definition (e.g. "I want us to be equals") carries implicit presuppositions:

(1) Speaker has standing to characterize the relationship

(2) Speaker's characterization is relevant and warranted

(3) Receiver should accept speaker's framing

These presuppositions encode dominance in R(M). Formally: f(R(M)) > 0 (dominance signal) despite C(M) asserting symmetry (leveling content).

The meta-level framing overrides the content-level claim.

Step 2: Receiver extracts dominance from message structure (from A5).

By A5's relational dynamics, agent j extracts meta-signals from R(M) to update π_ij. From the meta-communicative framing:

$$f(R(M)) > 0$$

Therefore:

$$\pi\_ij(t+1) = \pi\_ij(t) + f(R(M)) + \eta\_t > \pi\_ij(t)$$

Result: $|\pi\_ij(t+1)| > \varepsilon$, crossing into complementarity rather than retreating toward symmetry—opposite to i's intention.

Step 3: Belief update makes dominance explicit (from A1). Agent j observes M and updates beliefs about both content and relationship:

From C(M): Bel_j includes ("i claims to want equality", 0.8)

From R(M) (the meta-act structure): Bel_j includes ("i is defining our relationship", 0.7), ("i assumes right to characterize us", 0.6)

The belief update in Step 3 does not cause the dominance signal—it *recognizes* the dominance already present in R(M) per Step 1. Agent j's recognition reflects the pragmatic implication: if i truly viewed us as equals, i would not unilaterally declare our equality. The declaration itself presumes asymmetric authority to define terms.

Step 4: Paradoxical feedback under repetition. If agent i perceives that j still views relationship as asymmetric, i may intensify meta-communicative efforts: "I really mean it—I'm not trying to dominate." Each repetition:

- Reinforces i's unilateral right to define relationship (meta-dominance)
- Increases j's belief: ("i is protesting too much", 0.8), ("i is controlling", 0.85)
- Further amplifies $f(R(M)) > 0$

After k repetitions: $\pi\_ij(t+k) \approx \pi\_ij(t) + k \cdot f(R(M))$

The more i explicitly denies dominance, the more π_ij grows.

Why self-correction fails (the double bind). Agent i faces a pragmatic trap:

(1) Option A: Continue meta-communicating symmetry → increases dominance signal

(2) Option B: Stop meta-communicating → interpreted as giving up, possibly confirming dominance

(3) Option C (escape): Demonstrate symmetry through behavioral leveling (turn-sharing, soliciting input, yielding floor) without explicit characterization Only Option C breaks the paradox, because behavioral signals in R(M) lack the meta-framing problem: actions speak without claiming authority to define. Clinical/applied note. This paradox appears across contexts:

(4) Therapy: Therapist saying "This is a safe space where we're equals" may inadvertently assert therapeutic authority

(5) Management: Boss declaring "My door is always open—we're all on the same team" signals hierarchy while claiming to erase it

(6) Intimate relationships: "Don't worry, I'm not trying to control you" often precipitates precisely the control anxiety it aims to alleviate

(7) Pedagogy: Teacher proclaiming "I'm learning from you too" while grading student work

The formalization reveals why these well-intentioned statements backfire: A2 ensures that the meta-communicative act itself carries R(M) that contradicts C(M). Intervention requires shifting from explicit relationship definition to implicit behavioral demonstration.

Empirical predictions from T7

Prediction 1 (Paradoxical amplification): In dyads near complementarity threshold ($\varepsilon - 0.1 < |\pi\_ij| < \varepsilon + 0.1$), explicit symmetry assertions should increase $|\pi\_ij|$ by measurable amount ($\Delta\pi > 0.2$) within one exchange, compared to no-statement control condition. Effect size should correlate with meta-communicative explicitness (e.g. "I value equality" > "let's collaborate" > behavioral turn-sharing).

Prediction 2 (Repetition effect): Agents who make 3+ symmetry assertions should show steeper π_ij trajectory toward complementarity than those making 1–2 assertions or none. Relationship between assertion count and π_ij should be monotonic increasing, distinguishing this from random fluctuation.

Prediction 3 (Behavioral escape): Dyads instructed to demonstrate symmetry through behavior (equal turn time, collaborative decisions) without verbal framing should show π_ij decrease toward 0, while dyads instructed to verbally affirm symmetry should show π_ij increase. This dissociation confirms that the paradox operates at meta-communicative level, not content level.

Prediction 4 (Receiver attribution): When shown video of agent making symmetry assertion, third-party coders should rate agent as "trying to control the relationship definition" (dominance attribution) more than when shown same agent engaging in turn-sharing behavior without verbal framing. This validates that R(M) of meta-communicative act signals dominance independent of observer role.

Operationalization. Test in lab using confederate paradigm: manipulate whether confederate makes 0, 1, or 3 explicit symmetry statements ("I want us to work as equals") while holding behavioral turn-taking constant. Measure participant's continuous ratings of "who has more influence in this interaction" as proxy for π_ij. Predict inverted-U or monotonic increase with statement count. Control condition: confederate demonstrates turn equity without verbal framing—predict π_ij near 0. ∎

Table 2 provides a bridge from theory to practice. It maps each axiom and theorem to concrete dependent measures, experimental manipulations, and predicted observable effects, with brief notes on suitable analyses. The goal is to make our formal claims operational: readers can see exactly what to measure (e.g. belief change, relational coding, segmentation), what to vary (e.g. code choice, latency, repair cost), and what patterns should appear if the axioms hold (e.g. entropy reductions, reciprocal blame, stabilization of π). It also serves as a pre-registration scaffold for empirical tests and a replication checklist, thereby reducing the researchers' degrees of freedom by specifying variables and expected directions *a priori*. Finally, the table clarifies how the levels in our framework connect information flow, meta-content, coding, attribution, and social dynamics, so reviewers and practitioners can target interventions at the mechanism most likely to be at fault.

Discussion

Watzlawick *et al.* (1967) offered field-shaping propositions, but their force was primarily heuristic: powerful for reframing practice, yet difficult to test. Our contribution is to restate those insights as axioms with explicit scope and force, i.e. level-separating information flow (A1), meta-content (A2), code choice (A3), subjective punctuation (A4), and relational dynamics (A5). In this form, the claims become constraints on admissible models rather than explanatory slogans, enabling proofs, ablations, and falsification (cf. Bateson, 1972; Scott, 1997).

We argue that this formalization offers an advance. Specifically, three differences matter. First, operational precision: each primitive (e.g. $B\_i(t)$, $M\_\{i\to j\}(t)$, R(M), p_i, $\pi\_ij(t)$) can be observed, estimated, or manipulated. Second, independence and modularity: the axioms are weak enough to admit counter-models, so independence can be tested by selectively violating one while holding others fixed. Third, derivable consequences: our T1–T3, T6– (and T4–T5) are not restatements but entailments; e.g. that intentional silence is communicative under co-

Table 2. Operational map from axioms to observables in socio-technical interaction. Shaded rows indicate theorems demonstrating pathological rather than normative interaction patterns (π_ij = relational state; λ = leakage; Θ = timing norms; ε = symmetry/complementarity threshold; $\bar{H}$ = mean entropy

Item	Construct (level)	Candidate measures (DV)	Experimental manipulation (IV)	Predicted observable effect	Notes / analysis
A1: Unavoidability	Belief update (information)	Confidence ratings on key propositions (pre/post); Bayesian surprise; proposition addition/removal above threshold θ; response-time change	Presence/absence of observable behavior (incl. silence); visibility/noise; co-presence vs. no-observation	Observable B_$i(t)$ produces measurable change in Bel_j (t+1): confidence shifts $\|\Delta c\| > \varepsilon$, or new propositions entering above θ; larger effects under clearer observation	Pre-post paired t-tests on confidence scores; Bayesian surprise metrics; mixed-effects models with subject random intercepts
A2: Meta-inevitability	Relational content R(M)	Stance coding (dominance/affiliation); rights–duties attributions; perceived intent	Hold C(M) constant; vary relational cues (politeness, address terms, prosody)	Non-zero R(M) ratings even with "neutral" content; stronger effects with richer cues	Inter-rater reliability; ordinal models
A3: Multi-code realizability	Code/medium and variance	Posterior interpretive entropy; between-rater variance; misinterpretation rate	Same C(M), encode: digital vs. analog-rich vs. mixed	Lower $\bar{H}$ for mixed/analog-rich vs. digital encodings, conditional on channel/leakage (see A10)	Information-theoretic comparisons of H/h across conditions; standard homogeneity tests (e.g. Levene, 1960; Brown and Forsythe, 1974) if using parametric models on per-subject entropy scores
A4: Subjective punctuation	Segmentation p_i and attribution	Episode boundary placement; "who started/escalated" judgments	Ambiguous onset/offset; timeline reconstruction prompts	Divergent $p\_i \Rightarrow$ reciprocal blame without changing C(M)	Cross-tab of attributions; κ for boundary agreement
A5: Relational patterning	Dyadic state $\pi\_ij(t)$	Continuous dominance/affiliation ratings; turn-share; interruption rate	Repeated meta-signals (leveling vs. dominance); feedback constraints	Thresholding: $\|\pi_{ij}\|$ crosses $\varepsilon \rightarrow$ stable symmetry ($\leq\varepsilon$) or complementarity ($>\varepsilon$)	
A6: Turn-taking and timing	Floor rights / norms Θ	Latency, overlap, floor acquisition rate; perceived dominance	Enforce short vs. long latencies; allow/forbid overlap	Short latencies ↑ dominance attribution; miscoordination ↑ R(M) variance	Mixed-effects; latency as predictor of dominance

(*continued*)

Table 2. Continued

Item	Construct (level)	Candidate measures (DV)	Experimental manipulation (IV)	Predicted observable effect	Notes / analysis
A7: Repairability	Error handling	Repair initiation/resolution rates; uncertainty reduction	Vary repair cost (time/ penalty); block vs. invite repair	Higher cost ↓ repair → ↑ misattribution and escalation	Mediation: repair → uncertainty ↓ → blame ↓
A8: Common ground	Grounding / presupposition	Grounding moves; acceptance vs. challenge rates	Introduce novel terms with/ without grounding	Missing grounding ↑ failure-to-uptake despite clear code	Logistic models for uptake
A9: Relevance/ Cooperation	Contribution design	Relevance ratings; perceived engagement/dominance	Insert off-goal, over-informative, or under-informative turns	Irrelevance shifts (R(M))): disengagement or dominance, context-dependent	Interaction terms: task goal × relevance
A10: Public vs private channels	Signaling vs. state; leakage (λ)	Prosodic/physio proxies; micro-expressions; perceived sincerity	Induce strategic encoding; vary leakage via channel constraints	Moderate (λ) ↓ interpretive entropy; very low/high (λ) ↑ uncertainty	Quadratic (inverted-U) fits for λ
T1: Communicativity of silence	Silence as B_$i(t)$	Belief Δ; inferred intent after silence	Respond vs. remain silent following prompt	Silence produces systematic Bel_j update	Compare to baseline noise; equivalence tests
T2: Punctuation asymmetry ⇒ blame	Divergent p_i	Mutual blame index; escalation ratings	Provide same stream, manipulate cut-points via instructions	Reciprocal blame emerges from punctuation differences alone	Preregistered contrast of blame symmetry
T3: Meta-signals stabilize patterns	π dynamics	Convergence to symmetric/ complementary regime	Repeated dominance vs. leveling meta-signals	Trajectories converge to $\lvert\pi_{ij}\rvert \le \varepsilon$ (symmetric) or $> \varepsilon$ (complementary)	
T4: Early-anchor vulnerability	Anchoring × timing	Persistence of initial misinterpretation	Early misleading anchor; restrict correction windows (latency)	Anchors resist later corrections when outside (Θ)	Interaction: timing × correction; partial η^2

(*continued*)

Table 2. Continued

Item	Construct (level)	Candidate measures (DV)	Experimental manipulation (IV)	Predicted observable effect	Notes / analysis
T5: Credibility via code choice	Code × leakage	Entropy/order of certainty about R(M)	Cross code conditions, vary λ	Across code conditions, compare group-mean posterior interpretive entropy $\bar{H}$; expect lower $\bar{H}$ for mixed/analog-rich encodings at moderate leakage, with higher uncertainty at very low or very high leakage (inverted-U in λ). This "minimum" is a descriptive characterization of uncertainty under the model, not a normative optimization target; lower entropy need not imply better relational outcomes	Comparative entropy ordering across conditions; avoid optimization language. Fit monotone segments by code and a quadratic (or GAM) curve for λ to test an inverted-U
T6: Symmetrical escalation	Punctuation divergence × reactivity	Escalation trajectory; mutual blame indices; behavioral intensity (volume, interruptions)	Manipulate *p_i* vs *p_j* alignment; vary repair availability; measure reactivity (α parameter)	Divergent *p_i*, *p_j* + no repair → increasing intensity; aligned *p_i*, *p_j* OR repair access → de-escalation	Time-series models; breakpoint analysis; mediation analysis (repair → intensity ↓) |
T7: Meta-communication paradox	Meta-communicative dominance × relational threshold	π_{ij} trajectory; dominance attribution ratings; behavioral vs verbal symmetry signals	Manipulate: confederate makes 0, 1, or 3 explicit symmetry statements ("I value equality") while holding behavioral turn-taking constant; measure near threshold (ε ± 0.1)	Explicit symmetry assertions increase $\lvert\pi_{ij}\rvert$ ($\Delta\pi > 0.2$); monotonic increase with statement count; behavioral demonstration (no verbal framing) decreases π_{ij} toward 0; third-party coders rate verbal claims as "trying to control definition" more than behavioral symmetry	Inverted-U or monotonic fits; third-party coding for "trying to control definition"; dissociation between verbal claims and behavioral effects; confederate paradigm

presence (A1), that punctuation asymmetry rationally yields reciprocal blame (A4), and that repeated meta-signals stabilize symmetry or complementarity (A5).

Our proposed framework establishes links to micro-mechanisms. A6–A10 integrate classic conversation-analytic and pragmatic mechanisms into the axiomatics: turn-taking (Sacks *et al*., 1974), repair (Schegloff *et al*., 1977), common ground (Clark, 1996), and defeasible cooperation/relevance (Grice, 1975); plus an explicit public–private channel separation (A10). This enlarges the explanatory scope while keeping level separation: timing affects attribution (A6↔A5); repair modulates belief and relation (A7↔A1/A2); grounding constrains uptake (A8↔A2); and strategic signaling/leakage (A10) refines A3's code-variance story, making these interactional mechanisms available to observer-dependent, second-order modeling.

Our axioms deliberately sit at the intersection of formal belief-change, dynamical systems, signaling, and micro-interaction theory. For belief updates (A1–A2), dynamic epistemic logic (DEL) provides a semantics in which observations ("epistemic actions") transform agents' information states; public-announcement and action-model frameworks can be instantiated to our Bel_$j(t)$ operator and used to make scope and common-ground assumptions explicit (van Ditmarsch, van der Hoek and Kooi, 2007). To operationalize "communication happened," we can pair DEL with Bayesian surprise or related information-theoretic measures as empirical proxies for belief change (Itti and Baldi, 2009). For the relational state $\pi\_ij(t)$ (*A5*), state-space/dynamical models offer estimators and hypothesis tests for thresholding, stabilization, and perturbation, e.g. linear-Gaussian or switching state-space models with inputs derived from meta-signals R(M) (Durbin and Koopman, 2012; Shumway and Stoffer, 2017).

On code choice and leakage (A3, A10), our account aligns with signaling theory, where equilibria depend on how costly and diagnostic different channels are; mixed digital/analog encodings with partial leakage can be framed as separating vs. pooling regimes that order posterior uncertainty about R(M) (Spence, 1973). Game-theoretic pragmatics (Franke and Jäger, 2016) provides explicit models of how code choice affects interpretive variance through iterated best-response reasoning, directly instantiating the variance-ordering predictions in our A3 and T5. Similarly, relevance theory (Sperber and Wilson, 1986/1995) offers processing-cost models that predict when analog-rich codes (requiring more inference) versus digital codes (explicit but context-free) minimize interpretive entropy—a prediction central to our T5.

Finally, A6–A9 connect the axiomatics to established micro-mechanisms: turn-taking as temporal allocation of floor rights (Sacks *et al*., 1974), repair as the system's native error-correction (Schegloff *et al*., 1977), with Heritage (1984) providing the ethnomethodological foundation linking these sequential structures to participants' accountability practices, common ground as the substrate of uptake (Clark, 1996), and defeasible cooperativity/relevance as the pragmatic prior on contribution design (Grice, 1975). Together, these literatures give our level-separated axioms off-the-shelf semantics, estimators, and experimental levers.

Importantly, the proposed framework licenses concrete metrics: (1) belief-update proxies for A1 (e.g. pre/post inference tests, Bayesian surprise); (2) relational coding for A2 (annotating stance, rights/obligations, affiliation/dominance cues), building on speech act theory and implicature frameworks developed in formal pragmatics (Levinson, 1983); (3) variance/entropy comparisons across digital vs. analog encodings for A3; (4) elicitation or manipulation of segmentation functions p_i for A4 (e.g. timeline reconstruction tasks); and (5) estimation of relational state $\pi\_ij(t)$ for A5 (e.g. dynamic ratings or state-space models). These measures are portable across settings, e.g. therapy, teamwork, human computer interaction, and training.

Empirical agenda: designs and predictions
We sketch below three families of studies:

(1) *Code-variance tests* (*A3+A10*): Hold C(M) constant; randomize code (digital, analog-rich, mixed) and leakage. Predict a partial order on posterior interpretive entropy, with mixed/analog-rich encodings predicting systematic changes in posterior interpretive entropy over R(M), with lower dispersion at moderate leakage and higher dispersion at very low or very high leakage. (T5).

(2) *Punctuation manipulation* (*A4*): Randomize segmentation prompts or introduce ambiguous onsets; predict reciprocal blame and divergent causal attributions without changing content (T2).

(3) *Relational dynamics* (*A5*): Repeated meta-signals over time; estimate $\pi\_ij(t)$ and test threshold dynamics for stabilization into symmetry or complementarity (T3). Pre-registering analyses allows adjudication among alternative update functions f(·).

Implications for observing systems
A second-order perspective requires that the distinctions used to describe interaction are themselves made by an observing system. Our axiom set specifies what such an observer must minimally register. This is compatible with communication-as-autopoiesis views, where communication reproduces itself through selections made by observers (Luhmann, 1995). A1–A3 tell the observer that any co-present behavior is potentially belief-changing, that every communicative uptake includes a relational/meta attribution, and that code choice constrains the spread of possible interpretations. A4 tells the observer that segmentation and causal attributions are agent-relative; different observers will "see" different initiations and escalations in the same stream. A5 tells the observer that relational stance is not an external label but a state variable, $\pi ij(t)$, that is updated by meta-signals over time. In other words, the axioms define the observation task.

In this non-ontological reading, conducts are taken as action schemata and meaning is the outcome of operative distinctions by participants. Our formalism therefore treats *R*(M) and *C*(M) as attributed dimensions of uptake, not as properties of the conduct, and uses entropy strictly to summarize dispersion of readings conditional on a specific I_j, p_j.

The additional axioms (A6–A10) extend this task from "what is communicated" to "how interaction is organized." An observing system that ignores timing (A6), repair (A7), or common ground (A8) will systematically misclassify episodes as failed or hostile when, in fact, they are locally coherent. Likewise, without an explicit place for cooperation/relevance defaults (A9) and for public–private leakage (A10), the observer cannot explain why two utterances with identical content and code produce different relational updates. This is precisely the situation second-order cybernetics warned about: explanations that omit the observer's own distinctions make interaction look more paradoxical than it is (von Foerster, 1979; Glanville, 2004).

Seen this way, the present formalization offers a pragmatic companion to second-order cybernetics. It shows that the classic insights of Bateson and Watzlawick can be rendered as observer-dependent rules and still yield testable predictions (e.g. systematic changes in posterior interpretive entropy under different code/leakage regimes). It also shows that including reflexivity and meta-communication is not an obstacle to formal treatment but a requirement for modeling real socio-technical interaction (Pask, 1975; Brier, 2008).

Relative to recent work in Kybernetes, our contribution is not to restate observer-dependence in general terms, but to specify a minimal set of interactional distinctions that an observing system must track if pragmatic communication is to be made testable. In this sense, the paper complements recent discussions of second-order observation, operational epistemology and social systems theory (Laursen *et al.*, 2022; Roth and Sales, 2025) by

shifting the focus from abstract observer-inclusion to concrete observer-dependent segmentation, meta-signal uptake and relational state dynamics.

Applications and interventions

For practitioners, the calculus points to where to intervene. If breakdowns stem from A4/T2/T6 (punctuation, reciprocal blame, or symmetrical escalation), training should synchronize segmentation before debating content; a principle supported by formal mediation analysis (Raiffa, 1982), which emphasizes shared problem framing as a prerequisite to negotiation. If they arise from A3/A10/T5 (code and leakage), choose encodings that appropriately constrain interpretive dispersion in relation to the task and relational setting. If A6/A7 are implicated (timing/repair), lower repair costs and adjust latency norms. If A5/T3 dominates, design meta-signal counter-patterns (i.e. leveling moves) to push $\pi\_ij(t)$ back toward symmetry. These intervention points align with cooperative equilibrium strategies formalized in evolutionary game theory (Axelrod, 1984), where reducing noise, enabling signaling and establishing reciprocity norms prevent defection cascades — structural analogues to our escalation and stabilization theorems.

T6 on symmetrical escalation offers specific leverage points. If assessment reveals divergent punctuation (e.g. in couples therapy, each partner recounts the conflict with different initiating events), the intervention targets A4: use timeline reconstruction to synchronize p_i and p_j before addressing content. If punctuation is aligned but escalation persists, the issue may be suppressed repair (A7): lower the cost of clarification by establishing "meta" time-outs where either party can request process discussion without penalty. If reactivity is high ($\alpha > 1$), cognitive interventions to reduce amplification—reframing the other's behavior as reactive rather than initiatory, can dampen the feedback loop.

The Structured Democratic Dialogue methodology developed for Cyprus peace-building (Broome, 1998, 2002; Broome and Anastasiou, 2011; AUTHORS) exemplifies systematic A4 intervention at scale.

A practical payoff of level-separated axioms is in human–AI or multi-agent interaction. In this sense, A4, A6–A8 and T5 do not merely suggest generic "better communication," but specify which formal mechanisms underlie particular classes of socio-technical breakdown and therefore what kind of repair or redesign is required. A4 (subjective punctuation) and A10 (public–private channels/leakage) describe exactly the loci where human and artificial agents misalign about who initiated an episode, what is face-threatening, or whether a silence "counts" as a move. Implementing A6–A8 in dialogue managers (timing, repair, grounding) would reduce such misalignments in socio-technical systems. Recent work on common ground in HCI (Brennan, 1998; Skantze, 2021) shows that conversational AI systems systematically underperform on grounding moves μ specified in our A8, leading to breakdowns even when utterance generation is fluent. Similarly, formal presupposition models from Stalnaker (2002) reveal that AI dialogue managers often violate the $CG(t) \subseteq$ accessibility constraint, producing contextually inappropriate responses despite semantic well-formedness. Our framework provides diagnostic leverage: failures traced to A6 (timing) require different fixes than those traced to A8 (grounding) or A4 (punctuation), enabling targeted rather than ad-hoc improvements.

What formalization cannot capture

Our axiomatization deliberately omits aspects of interaction central to lived experience but resistant to formal treatment: the felt sense of connection or alienation, the aesthetic qualities of conversation, the ethical weight of speech acts, and the ways context shapes meaning beyond cognitive belief update. These aspects become increasing important, especially today when many structured interactions take place over virtual channels (for example, SDD has evolved to allow hybrid virtual/face-to-face as well as synchronous/asynchronous interactions: AUTHORS). These are not oversights but acknowledgments that

formalization is a tool for specific purposes—enabling measurement, deriving predictions, designing interventions—rather than a complete account of human communication. We echo Bateson's (1972) caution that "the map is not the territory," offering our framework as a map useful for navigation while recognizing the territory's irreducible complexity.

Limitations and scope

Our framework is intentionally minimal. It abstracts from rich sociocultural scaffolding, higher-order beliefs about beliefs, and power structures beyond dyads. It also presumes observability sufficient for belief update (A1), which may fail under noise or deception. These choices keep axioms testable while leaving room to extend with dynamic epistemic logic for higher-order reasoning and state-space or reinforcement-learning formalisms for strategic interaction.

Our formalization of belief states as graded commitments (Bel_$j(t) = \{(p, c)\}$) represents one tractable operationalization among several possible choices. We selected this hybrid approach—between fully propositional (all-or-nothing) and fully probabilistic (Bayesian)—to balance formal rigor with empirical feasibility and clinical applicability. Alternative formalizations are viable: a strict Bayesian treatment would represent Bel_$j(t)$ as probability distributions over hypothesis spaces and specify likelihood functions for each behavior type, enabling principled surprise calculations but requiring parameter estimates we do not yet possess; a purely propositional approach would simplify to set operations (belief revision logic; Alchourrón *et al*., 1985) but would struggle to capture the graded nature of relational interpretations central to A2 and A5. Future work should test whether key theorems hold under alternative belief representations, and whether the choice of formalism affects the predictive validity of derived interventions. Our framework is designed to be robust to this choice: the core axioms constrain that beliefs must update (A1) and what dimensions shift (content vs. relational), not the precise mathematical form of the update operator U_j.

Conclusion

By converting Watzlawick's program into a testable axiomatics, we keep the spirit of "pragmatics" while delivering a toolbox for proof, measurement, and intervention. The point is not to replace clinical wisdom with equations but to stabilize cumulative science: precise enough to fail, modular enough to grow, and practical enough to help real groups talk—and think—better (Bateson, 1972; Watzlawick *et al*., 1967; Scott, 1997). By demonstrating that even paradoxical, reflexive interaction patterns yield to systematic analysis, we hope to encourage similar formalization efforts across the social sciences—not to replace humanistic insight, but to create cumulative knowledge that can be taught, tested, and translated into practice.

Note

1. In dyadic interaction (i, j), N typically refers to one receiver per message (H_bar = H_j for M_$\{i \to j\}$), or to multiple messages across the dyad. For empirical tests (e.g. T5), N may refer to third-party coders using a predetermined R(M) scheme; cross-coder entropy measures interpretive ambiguity given the coding protocol, not agent-level variance.

References

Alchourrón, C.E., Gärdenfors, P. and Makinson, D. (1985), "On the logic of theory change: partial meet contraction and revision functions", *Journal of Symbolic Logic*, Vol. 50 No. 2, pp. 510-530, doi: 10.2307/2274239.

Axelrod, R. (1984), *The Evolution of Cooperation*, Basic Books, available at: https://ee.stanford.edu/%7Ehellman/Breakthrough/book/pdfs/axelrod.pdf

Bateson, G. (1972), *Steps to an Ecology of Mind*, University of Chicago Press, Chicago.

Brennan, S.E. (1998), "The grounding problem in conversations with and through computers", in Fussell, S.R. and Kreuz, R.J. (Eds), *Social and Cognitive Approaches to Interpersonal Communication*, Lawrence Erlbaum Associates, pp. 201-225.

Brier, S. (2008), *Cybersemiotics: Why Information is Not Enough*, University of Toronto Press, Toronto.

Broome, B.J. (1998), "Overview of conflict resolution activities in Cyprus: their contribution to the peace process", *Cyprus Review*, Vol. 10 No. 1, pp. 47-66, available at: https://cyprusreview.org/index.php/cr/article/view/490

Broome, B.J. (2002), "Participatory planning and design in a protracted conflict situation: applications with citizen peace-building groups in Cyprus", *Systems Research and Behavioral Science: The Official Journal of the International Federation for Systems Research*, Vol. 19 No. 4, pp. 313-321, doi: 10.1002/sres.434.

Broome, B.J. and Anastasiou, H. (2011), "Communication across the divide in the Cyprus conflict", in *Handbook of Ethnic Conflict: International Perspectives*, Springer US, Boston, MA, pp. 293-324, doi: 10.1007/978-1-4614-0448-4_12.

Brown, M.B. and Forsythe, A.B. (1974), "Robust tests for the equality of variances", *Journal of the American Statistical Association*, Vol. 69 No. 346, pp. 364-367, doi: 10.1080/01621459.1974.10482955.

Clark, H.H. (1996), *Using Language*, Cambridge University Press, Cambridge.

Clark, H.H. and Brennan, S.E. (1991), "Grounding in communication", in Resnick, L.B., Levine, J.M. and Teasley, S.D. (Eds), *Perspectives on Socially Shared Cognition*, American Psychological Association, pp. 127-149, doi: 10.1037/10096-006.

Cover, T.M. and Thomas, J.A. (2006), *Elements of Information Theory*, 2nd ed., Wiley, doi: 10.1002/047174882X.

del Rio, C.M. (2012), "Book review: pragmatics of human communication: a study of interactional patterns, pathologies and paradoxes", *The Family Journal*, Vol. 20 No. 3, pp. 341-343, doi: 10.1177/1066480712449802.

Durbin, J. and Koopman, S.J. (2012), *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press, doi: 10.1093/acprof:oso/9780199641178.001.0001.

Franke, M. and Jäger, G. (2016), "Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics", *Zeitschrift für Sprachwissenschaft*, Vol. 35 No. 1, pp. 3-44, doi: 10.1515/zfs-2016-0002.

Glanville, R. (2002), "Second-order cybernetics", in Parra-Luna, F. (Ed.), *Systems Science and Cybernetics*, Kluwer, pp. 59-74.

Glanville, R. (2004), "The purpose of second-order cybernetics", *Kybernetes*, Vol. 33 Nos 9/10, pp. 1379-1386, doi: 10.1108/03684920410556016.

Godat, D. and Czerny, E.J. (2021), "Communication today: were Watzlawick & Co. wrong?", *Journal of Solution Focused Practices*, Vol. 5 No. 2, p. 10, available at: https://digitalscholarship.unlv.edu/journalsfp/vol5/iss2/10.

Grice, H.P. (1975), "Logic and conversation", in Cole, P. and Morgan, J.L. (Eds), *Syntax and Semantics*, Academic Press, Vol. 3, pp. 41-58.

Haley, J. (1963), *Strategies of Psychotherapy*, Grune & Stratton, available at: https://psycnet.apa.org/doi/10.1037/14324-000

Haley, J. (1987), *Problem-Solving Therapy*, 2nd ed., Jossey-Bass, available at: https://psycnet.apa.org/record/1987-98523-000

Heritage, J. (1984), *Garfinkel and Ethnomethodology*, Polity Press, Cambridge.

Itti, L. and Baldi, P. (2009), "Bayesian surprise attracts human attention", *Vision Research*, Vol. 49 No. 10, pp. 1295-1306, doi: 10.1016/j.visres.2008.09.007.

Laursen, K.B., Harste, G. and Roth, S. (2022), "Moral communication observed with social systems theory. An introduction", *Kybernetes*, Vol. 51 No. 5, pp. 1653-1665, doi: 10.1108/K-01-2022-0059.

Levene, H. (1960), "Robust tests for equality of variances", in Olkin, I. (Ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford University Press, pp. 278-292.

Levinson, S.C. (1983), *Pragmatics*, Cambridge University Press, Cambridge, doi: 10.1017/CBO9780511813313.

Luhmann, N. (1995), *Social Systems*, Stanford University Press, Stanford, Original work published 1984.

Lutterer, W. (2007), "The two beginnings of communication theory", *Kybernetes*, Vol. 36 Nos 7/8, pp. 1022-1025, doi: 10.1108/03684920710777793.

MacKay, D.M. (1969), *Information, Mechanism and Meaning*, MIT Press, Cambridge, Massachusetts.

Osejo-Bucheli, C. (2025), "Towards a unified ontology of cybernetics: bridging mechanology and CAS for cooperative societies through thematic synthesis", *Kybernetes*, Vol. 54 No. 12, pp. 7061-7082, doi: 10.1108/K-03-2024-0553.

Pask, G. (1975), *Conversation, Cognition and Learning: a Cybernetic Theory and Methodology*, Elsevier, Amsterdam.

Pask, G. (1976), "Conversation theory: applications in education and epistemology", available at: https://api.semanticscholar.org/CorpusID:142493376

Raiffa, H. (1982), *The Art and Science of Negotiation*, Harvard University Press, Cambridge, Massachusetts.

Roth, S. and Sales, A. (2025), "Editorial: cybernetics and systems theories for the 21st century. Introducing the new aims and scope of kybernetes", *Kybernetes*, Vol. 54 No. 8, pp. 4071-4077, doi: 10.1108/K-07-2025-330.

Ruesch, J. and Bateson, G. (1951), *Communication: the Social Matrix of Psychiatry*, W. W. Norton, New York.

Sacks, H., Schegloff, E.A. and Jefferson, G. (1974), "A simplest systematics for the organization of turn-taking for conversation", *Language*, Vol. 50 No. 4, pp. 696-735, doi: 10.2307/412243.

Saratxaga Arregi, A. (2025), "Heinz von Foerster's operational epistemology: orientation for insight into complexity", *Kybernetes*, Vol. 54 No. 10, pp. 5891-5910, doi: 10.1108/K-10-2023-2116.

Schegloff, E.A., Jefferson, G. and Sacks, H. (1977), "The preference for self-correction in the organization of repair in conversation", *Language*, Vol. 53 No. 2, pp. 361-382, doi: 10.1353/lan.1977.0041.

Scott, B. (1997), "Inadvertent pathologies of communication in human systems", *Kybernetes*, Vol. 26 Nos 6/7, pp. 824-836, doi: 10.1108/03684929710170021.

Shannon, C.E. (1948), "A mathematical theory of communication", *Bell System Technical Journal*, Vol. 27 Nos 3-4, pp. 379-423, doi: 10.1002/j.1538-7305.1948.tb01338.x.

Shumway, R.H. and Stoffer, D.S. (2017), *Time Series Analysis and its Applications: with R Examples*, 4th ed., Springer, doi: 10.1007/0-387-36276-2.

Skantze, G. (2021), "Turn-taking in conversational systems and human-robot interaction: a review", *Computer Speech and Language*, Vol. 67, 101178, doi: 10.1016/j.csl.2020.101178.

Spence, M. (1973), "Job market signaling", *Quarterly Journal of Economics*, Vol. 87 No. 3, pp. 355-374, doi: 10.2307/1882010.

Sperber, D. and Wilson, D. (1986/1995), *Relevance: Communication and Cognition*, 2nd ed., Blackwell, Oxford. Original work published 1986.

Stalnaker, R. (2002), "Common ground", *Linguistics and Philosophy*, Vol. 25 Nos 5-6, pp. 701-721, doi: 10.1023/A:1020867916902.

Temizel, E. (2025), "ARCHITRAINER: building interaction at the intersection of architecture, cybernetics, psychology and technology", *Kybernetes*, Vol. 55 No. 6, pp. 2762-2780, doi: 10.1108/K-08-2024-2170.

van Ditmarsch, H., van der Hoek, W. and Kooi, B. (2007), *Dynamic Epistemic Logic*, Springer, Dordrecht.

von Foerster, H. (1974), *Cybernetics of Cybernetics (BCL Report 73.38)*, University of Illinois, Biological Computer Laboratory, Urbana, Illinois.

von Foerster, H. (1979), "Cybernetics of cybernetics", in Krippendorff, K. (Ed.), *Communication and Control in Society*, Gordon & Breach, pp. 5-8.

Watzlawick, P. (1984), *The Invented Reality: How do we Know what we Believe we Know?*, W. W. Norton, New York.

Watzlawick, P., Beavin Bavelas, J. and Jackson, D.D. (1967), *Pragmatics of Human Communication: a Study of Interactional Patterns, Pathologies, and Paradoxes*, W. W. Norton, New York.

Watzlawick, P., Weakland, J. and Fisch, R. (1974), *Change: Principles of Problem Formation and Problem Resolution*, W. W. Norton, New York.

Whitehead, A.N. and Russell, B. (1910-1913), *Principia Mathematica*, Cambridge University Press, Amsterdam, Vol. 1-3.

Further reading

Bateson, G. (1935), "Culture contact and schismogenesis", *Man*, Vol. 35, pp. 178-183, doi: 10.2307/2789408.

Russell, B. (1908), "Mathematical logic as based on the theory of types", *American Journal of Mathematics*, Vol. 30 No. 3, pp. 222-262, doi: 10.2307/2369948.

Corresponding author

Yiannis Laouris can be contacted at: laouris@cnti.org.cy