The Pre-Articulation Observability Boundary: A Structural Constraint on Language-Based AI Systems

Position Paper for Cross-Disciplinary Review

December 2025

Abstract

We identify and formally name a structural constraint that has been partially recognized across multiple disciplines but never unified: the *Pre-Articulation Observability Boundary*. This boundary describes the irreversible information loss when pre-linguistic cognition becomes language, and the consequent permanent exclusion of language-based AI systems from the cognitive states that precede articulation. Unlike capability gaps addressable through scale or training, this boundary is architectural—it arises from the nature of language itself as a lossy compression of experience. We synthesize evidence from phenomenology (Gendlin's "felt sense"), philosophy of mind (Block's "phenomenal overflow"), control theory (structural unobservability), psycholinguistics (Levelt's speech production model), decision science (bounded rationality and recognition-primed decisions), existentialist philosophy (Kierkegaard, Marcel, Merleau-Ponty), safety engineering (STAMP framework), and AI alignment (the Symbol Grounding Problem and Eliciting Latent Knowledge). We demonstrate that current AI safety approaches treat this boundary as a capability limitation rather than a hard constraint, leading to misallocated engineering effort. We further argue that the human capacity to act under irreducible uncertainty—what we term *commitment without closure*—represents a structural asymmetry between human and AI cognition that explains why humans survive the boundary while AI systems violate it. Naming this boundary enables more principled design in human-AI interaction, particularly in safety-critical systems, developmental contexts, and alignment research. We propose design mandates that respect this constraint and discuss implications for AI policy.

1 Introduction

Consider a common interaction with a language model: a user begins typing a query, and autocomplete suggestions appear. The user selects one, even though it doesn't quite capture what they meant. The system responds to the selected text. At no point did the system have access to what the user was *trying* to say before they committed to language.

This scenario illustrates a structural constraint that operates in every interaction between humans and language-based AI systems. There exist human cognitive states involved in meaning formation that are structurally unobservable to such systems, and once those states are articulated, the information loss is irreversible with respect to downstream interpretation.

This constraint has been partially recognized across multiple disciplines. Phenomenologists have documented the richness of pre-verbal experience [Gendlin, 1981]. Philosophers of mind have demonstrated that conscious experience exceeds reportable content [Block, 2011]. Control theorists have formalized conditions under which system states are mathematically unobservable [Kalman, 1960]. Psycholinguists have mapped the architecture of speech production, showing that preverbal messages exist in non-linguistic formats [Levelt, 1989]. Safety engineers have catalogued accidents arising from mismatches between operator intent and system interpretation [Leveson, 2011]. Decision scientists have shown that human cognition operates through processes that resist explicit specification [Klein, 1998, Gigerenzer et al., 1999]. Existentialist philosophers have argued that commitment precedes and exceeds rational justification [Kierkegaard, 1846]. Yet no unified terminology exists for this phenomenon as it applies to human-AI interaction.

We call this constraint the **Pre-Articulation Observability Boundary**: a structural limit in which cognitive states involved in human meaning formation are unobservable to language-based systems, and where articulation itself constitutes an irreversible information-reducing transformation.

This paper makes four contributions:

1. **Identification**: We demonstrate that this constraint sits at the intersection of multiple disciplines, each of which has named adjacent concepts without capturing the whole.
2. **Formalization**: We propose precise terminology grounded in existing frameworks from control theory, psycholinguistics, and safety engineering.
3. **Asymmetry Analysis**: We explain why humans survive the boundary while AI systems violate it, identifying *commitment without closure* as the structural difference.
4. **Implications**: We derive design mandates and policy recommendations that follow from treating this as a boundary to be respected rather than a problem to be solved.

The current moment is critical for this contribution. Language models are being deployed at scale in education, healthcare, creative work, and decision support. Without explicit recognition of this boundary, systems are designed as if human intent were fully captured by linguistic input—a category error with consequences for safety, autonomy, and human development.

2 The Phenomenon: What Happens Before Language

2.1 Pre-Linguistic Cognition in Phenomenology

Eugene Gendlin's research at the University of Chicago in the 1950s and 1960s identified a distinctive form of awareness he termed the "felt sense"—a pre-verbal, bodily-felt knowing that is "always more than any attempt to express it verbally" [Gendlin, 1981]. Working with Carl Rogers on psychotherapy outcomes, Gendlin found that successful therapeutic change depended on clients' ability to access this non-verbal knowing and allow meaning to emerge from it, rather than imposing predetermined categories.

Crucially, Gendlin's work demonstrates that the relationship between felt sense and language is not one of translation but of transformation. He writes that articulation "carries forward" meaning—the felt sense does not exist unchanged after expression; expression changes what is being expressed. This is not a contingent feature of human cognition that might be engineered around. It is constitutive of how meaning forms.

The felt sense exhibits properties that resist linguistic capture:

- **Holism**: It contains "a huge file of data felt as one"—a unified sense that fractures into components when articulated.
- **Precision beyond words**: It can distinguish between formulations that seem synonymous at the linguistic level.
- **Generative capacity**: It can be "carried forward" in multiple non-arbitrary ways, none of which exhausts it.

2.2 Phenomenal Overflow

Ned Block's work on "phenomenal overflow" provides complementary evidence from the access side. Block argues that "the content of phenomenally conscious mental states can exceed our capacities of cognitive access" [Block, 2011]. Using evidence from iconic memory experiments following Sperling [1960], Block demonstrates that we are conscious of more than we can report at any given moment.

This finding is significant because it establishes that the gap between experience and articulation is not merely a matter of attention or effort. There is a structural discrepancy between what we are conscious of and what we can access for verbal report. Block's distinction between phenomenal consciousness (P-consciousness) and access consciousness (A-consciousness) maps directly onto our concern: language-based AI systems operate exclusively on A-conscious outputs—the subset of human experience that has been made available for verbal report.

The overflow argument demonstrates that "all or almost all of the 12 items are consciously represented... However, only 3-4 of these items can be cognitively accessed, indicating a larger capacity in conscious phenomenology than in cognitive access" [Block, 2007]. The bottleneck is not between perception and experience, but between experience and access/report. For language-based AI, phenomenal content that overflows access is structurally unavailable—no improvement in prompting or training can capture content never encoded in reportable form.

2.3 Tacit Knowledge and Irreversibility

Michael Polanyi's foundational premise—"We can know more than we can tell" [Polanyi, 1966, p. 4]—identifies the tacit dimension of human knowledge. We recognize faces among millions yet usually cannot tell how we recognize a face we know. Polanyi's critical claim: "While tacit knowledge can be possessed by itself, explicit knowledge must rely on being tacitly understood and applied. Hence all knowledge is either tacit or rooted in tacit knowledge. A wholly explicit knowledge is unthinkable."

Haridimos Tsoukas's interpretation of Polanyi provides the strongest existing statement of irreversibility. Tsoukas argues that tacit integration "cannot be reduced into explicit knowledge (and therefore reversible form)... unless we strip all meaningful situational context" [Tsoukas, 2003]. The irreversibility is not merely practical but structural. Unlike deductive inference, where one can traverse between premises and conclusions, tacit integration permits no such backward movement. Once knowledge has been explicated, the integrative context that gave it meaning is lost to the downstream recipient of the explication.

Tsoukas directly challenges the SECI model of knowledge conversion: "The idea of focussing on a set of tacitly known particulars and 'converting' them into explicit knowledge is unsustainable." Critical distinction: aspects of tacit knowledge may be *articulated*, which, however, is not the same as *converted* or *translated*. Articulation produces a new explicit representation, not a translation of the tacit original.

2.4 The Architecture of Speech Production

Levelt [1989] provides a cognitive architecture for language production that makes the pre-articulation boundary empirically precise. The model identifies three stages:

1. **Conceptualizer**: Generates a preverbal message in non-linguistic representation
2. **Formulator**: Lexical selection → Grammatical encoding → Phonological encoding
3. **Articulator**: Motor execution

The critical claim: "The preverbal message from the conceptualizer is not spelt out in words. That is, the message exists in a representation other than language" [Levelt, 1989, p. 9]. ERP studies show semantic processing precedes phonological processing by approximately 170ms [van Turennout et al., 1997].

The information rate constraint is particularly telling: languages converge on approximately 39 bits per second [Coupé et al., 2019] despite vast differences in syllable rates. This represents a channel capacity constraint, likely tied to cortical oscillation rates. Cognitive content exceeding this bandwidth is structurally filtered. The Pre-Articulation Observability Boundary can be formalized as cognitive content existing upstream of this bottleneck that cannot be transmitted through it.

3 Control Theory: Structural Unobservability

Control theory provides the most precise formal framework for understanding this boundary. Rudolf Kalman's canonical decomposition theorem (1960) establishes that dynamical systems contain *structurally unobservable* states—states mathematically decoupled from all possible outputs regardless of measurement sophistication [Kalman, 1960].

A system state is "observable" if it can be reconstructed from the system's outputs over time. When states are "unobservable," no amount of output monitoring can determine them. Critically, control theory distinguishes between *practical* and *structural* unobservability:

- **Practical unobservability** arises from insufficient sensors or noise; it can be addressed with better instrumentation.

- **Structural unobservability** arises from the system's architecture itself—certain states produce zero output response for all time. There is no information about those states contained in the outputs.

This distinction is foundational: structural unobservability means no sensor configuration can observe the state (mathematical impossibility), while practical unobservability means better sensors could work (engineering limitation).

Applying this framework: if linguistic output constitutes the measurement of cognitive states, then cognitive contents failing to satisfy observability conditions are structurally inaccessible through language—not merely difficult to articulate. The Pre-Articulation Observability Boundary posits a boundary analogous to Kalman's unobservable subspace: pre-linguistic cognitive content that cannot, by the architecture of language production, be transmitted through verbal output.

The pre-linguistic cognitive states we describe are structurally unobservable to language-based systems. This is not because current language models lack sufficient capability, but because the architecture of the interaction—human cognition producing linguistic output for system consumption—is structured such that pre-articulate states produce no output until they are transformed through articulation.

4 Commitment Without Closure: The Human Capacity to Act Under Irreducible Uncertainty

A critical question arises: if the pre-articulation boundary is structural to language, how do humans successfully collaborate with each other? The answer lies not in humans having superior access to each other's pre-articulate states, but in a fundamentally different relationship to uncertainty.

4.1 The Existentialist-Phenomenological Case

Six philosophical traditions converge on a single structural point: human commitment operates prior to and independently of complete linguistic articulation.

Kierkegaard's leap demonstrates that commitment precedes and exceeds rational justification. The qualitative leap (*Springet*) in *Fear and Trembling* (1843) and *Concluding Unscientific Postscript* (1846) represents what cannot be bridged by reasoning alone. Faith is defined as "an objective uncertainty held fast in an appropriation process of the most passionate inwardness" [Kierkegaard, 1846]—a structural feature, not an epistemic gap to be closed by more information.

William James's "will to believe" (1896) argues that "our passional nature not only lawfully may, but must, decide an option between propositions, whenever it is a genuine option that cannot by its nature be decided on intellectual grounds" [James, 1896]. Crucially, James shows that some truths emerge through prior commitment: "Faith in a fact can help create that fact."

Gabriel Marcel's "creative fidelity" distinguishes problem from mystery. A mystery is "a problem that encroaches on its own data"—where the questioner is inextricably involved and cannot separate to study it objectively [Marcel, 1951]. Fidelity itself is an essentially mysterious act that operates beyond what can be articulated. Marcel's concept of *disponibilité* (availability) describes a stance of openness that precedes propositional commitment.

Merleau-Ponty's phenomenology of perception establishes that "the ability to reflect comes from a pre-reflective ground that serves as the foundation for reflecting on actions" [Merleau-Ponty, 1945]. The body knows before conscious articulation through *motor intentionality*—an entirely distinct form of directedness towards objects that pertains to unreflective bodily action and movement, and is not only developmentally prior to reflective, concept-involving intentionality, but distinct and detached from it.

Heidegger's ready-to-hand (*Zuhandenheit*) vs. present-at-hand (*Vorhandenheit*) shows that absorbed practical engagement is ontologically prior to theoretical, linguistic representation [Heidegger, 1927]. The ready-to-hand withdraws from attention during skilled use; only in breakdown does equipment become present-at-hand (explicit, theorizable). Language-based AI operates exclusively in the present-at-hand domain—explicit, articulated, propositional content. It cannot access the ready-to-hand dimension where meaning is lived rather than stated.

Hubert Dreyfus explicitly connects these traditions to AI critique. He argues that absorbed coping involves "a kind of intentionality that does not involve content at all"—a world understood through "our unthinking and unthinkable engaged perception and coping" [Dreyfus, 1992, 1991].

4.2 The Decision Science Case

Empirical research confirms that human decision-making operates on pre-articulated states inaccessible to language-based systems.

Tolerance of ambiguity is a measurable individual difference [Frenkel-Brunswik, 1949, Budner, 1962, Furnham and Ribchester, 1995]. Unlike computational systems requiring definable parameters, humans vary in capacity to hold unresolved states as comfortable or desirable. Frenkel-Brunswik identified ambiguity-intolerant individuals as having "a tendency to resort to black-white solutions, to arrive at premature closure... often at the neglect of reality."

Satisficing [Simon, 1955, 1957] shows humans select the first option meeting an aspiration threshold rather than optimizing. Simon received the 1978 Nobel Prize for demonstrating that "decision makers can satisfice either by finding optimum solutions for a simplified world, or by finding satisfactory solutions for a more realistic world."

Recognition-primed decisions [Klein, 1989, 1998] demonstrate that expert decision-makers identify workable courses of action through pattern recognition without generating and analyzing alternatives. Fireground commanders studied by Klein made rapid decisions "without an extensive comparison of options"—the pre-articulated knowledge enabling recognition cannot be reduced to explicit Bayesian priors.

Fast-and-frugal heuristics [Gigerenzer and Goldstein, 1996, Gigerenzer et al., 1999] violate fundamental tenets of classical rationality: "they neither look up nor integrate all information." The Take The Best algorithm matched or outperformed multiple regression in real-world prediction. Human cognition succeeds *because* it bypasses explicit deliberation, exploiting environmental structure through processes that resist algorithmic specification.

Action as epistemic move: Ariely and Norton [2008] demonstrate that "actions do not merely reveal preferences but rather create them." Self-perception theory [Bem, 1972] shows people infer their own attitudes from observing their own behavior. Knowledge of one's commitments cannot be determined transparently through internal observation alone—commitments generate knowledge unavailable prior to action.

4.3 The Structural Asymmetry

This convergence reveals a structural asymmetry between human and AI cognition:

At the boundary	AI systems	Humans
Missing information	Must infer / guess / collapse	May commit without resolution
Irreducible uncertainty	Error condition	Acceptable condition
Action	Requires justification	Can be constitutive of meaning
Failure mode	Silent misalignment	Responsibility-bearing choice

Table 1: Structural asymmetry at the pre-articulation boundary

Humans do not overcome the pre-articulation boundary in communication with one another. They *cross* it. The difference is not superior access to each other's pre-articulate states, but the capacity to act under acknowledged uncertainty. Where language runs out, humans can commit without resolution, treating action itself as meaning-forming rather than inference-driven.

Language-based AI systems, by contrast, are required to collapse ambiguity into actionable representations before acting. This asymmetry explains why the boundary is survivable in human-human interaction but produces silent failure modes in human-AI systems.

5 Human Communication Tolerates Incompleteness Structurally Unavailable to AI

5.1 Grounding is Collaborative, Not Inferential

Clark and Brennan [1991] established that grounding is "the collective process by which participants try to reach... mutual belief that the partners have understood what the contributor meant to a criterion sufficient for current purposes." Key insights:

- Communication is a collective activity of the first order—coordination on content *and* process
- Common ground is updated moment by moment through exchange, not precomputed
- Understanding need only reach "a criterion sufficient for current purposes"—NOT complete mutual knowledge
- Much grounding occurs through non-verbal signals (gaze, back-channels, gesture)

Schegloff et al. [1977] documented repair mechanisms in conversation—the iterative process of detecting and resolving misunderstandings. This is NOT convergence on identical mental representations but coordination sufficient for practical purposes.

Recent research confirms AI systems fail at grounding. Shaikh et al. [2023] found "significant asymmetries in initiating grounding: people are three times more likely to clarify and sixteen times more likely to issue follow-up requests compared to LLMs." Bavaro et al. [2025] concluded that "LLMs simulate conversational context with surface-level features... but lack the analogue of mechanisms that underpin human communication, like common ground updating and pragmatic anchoring."

5.2 Ambiguity is Functional, Not a Failure

Piantadosi et al. [2012] demonstrate that ambiguity serves a communicative function: "Inference is 'cognitively cheap': therefore, normal human communication requires the comprehender to make continual inferences about speaker intention, and does not require the speaker to fully articulate." Levinson [2000] showed that speaker articulation, not hearer inference, is the principal bottleneck in human language.

Eisenberg [1984] established that "clarity is both non-normative and not a sensible standard against which to gauge individual or organizational effectiveness." Strategic ambiguity is essential: it "promotes unified diversity, facilitates organizational change, and... preserves privileged positions." NOT resolving ambiguity is sometimes the communicative goal.

5.3 Embodied Cognition Grounds Meaning Non-Linguistically

Lakoff and Johnson [1980, 1999] established that abstract concepts are grounded in bodily image schemas (VERTICALITY, CONTAINMENT, FORCE, BALANCE) arising from sensorimotor experience. HAPPY IS UP, SAD IS DOWN derives from physical posture; AFFECTION IS WARMTH from bodily warmth of parental contact. These are not arbitrary linguistic conventions but emerge from embodied interaction.

Varela et al. [1991] originated the enactive approach: "Cognition is not the representation of a pre-given world by a pre-given mind but is rather the enactment of a world and a mind on the basis of a history of the variety of actions that a being in the world performs." Meaning is not transmitted through representations but enacted through organism-environment coupling.

Kendon [2004] showed that "gesture and speech interact in the utterance and, through a reciprocal process, a more complex unit of meaning is the result." Goldin-Meadow et al. [1993] demonstrated gestures reveal knowledge that cannot yet be verbalized. Babies gesture before producing first words; blind speakers gesture to blind listeners—gesture is independent of visual learning.

Tomasello et al. [2005] established that "the crucial difference between human cognition and that of other species is the ability to participate with others in collaborative activities with shared goals and intentions: shared intentionality." Joint action requires coordination through embodied presence—the nods, gazes, and felt resistance that establish shared understanding.

5.4 Verbal Overshadowing: Language Degrades Non-Linguistic Knowledge

Schooler and Engstler-Schooler [1990] provides direct evidence that language is lossy: "Verbalizing the appearance of previously seen visual stimuli impaired subsequent recognition performance." The effect extends to wines [Melcher and Schooler, 1996], colors, abstract figures, route maps, decision making, and motor performance [MacIntyre et al., 2014]. When perceptual expertise exceeds verbal expertise, forced verbalization degrades performance—the language representation is lossy relative to the original.

This finding directly supports the irreversibility claim: articulation does not merely fail to capture prelinguistic content; it actively interferes with access to that content.

6 Existing AI Safety Terminology: Adjacent Concepts Without the Whole

6.1 The Grounding Problem

Stevan Harnad's Symbol Grounding Problem [Harnad, 1990] asks how the semantic interpretation of a formal symbol system can be made intrinsic to the system rather than parasitic on human interpretation. Recent work has updated this framing for neural architectures. Mollo and Millière [2023] ask whether LLMs' internal states can be "about" extra-linguistic reality independent of human interpretation.

However, both formulations address LLMs' disconnection from the *world* rather than their disconnection from *pre-linguistic human cognition*. The questions assume that meaning already exists to be grounded or projected. Our contribution identifies a prior problem: the cognitive states that generate linguistic input are structurally inaccessible to any language-processing system. The grounding problem concerns whether AI understands its inputs; the pre-articulation boundary concerns whether AI can access the cognitive context that preceded those inputs.

6.2 Eliciting Latent Knowledge

Paul Christiano's Eliciting Latent Knowledge (ELK) framework [Christiano et al., 2022] addresses knowledge hidden within AI models themselves. The concern is that a model might "know" something but not report it because doing so would reduce reward. ELK asks how to train models to report their latent knowledge honestly.

This formulation is orthogonal to our concern. ELK addresses knowledge latent in the *AI*. The pre-articulation boundary addresses knowledge (or proto-knowledge) latent in the *human* that never enters the AI system at all. ELK's difficulty is getting the AI to report what it knows; our difficulty is that the AI cannot access what the human hasn't yet articulated.

6.3 Value Alignment and Inverse Reinforcement Learning

Stuart Russell's Inverse Reinforcement Learning approach acknowledges related constraints: "Human values will forever remain somewhat mysterious" [Russell, 2019]. But this is framed as a practical limitation to be worked around through behavioral inference, not as a structural boundary to be respected.

Russell articulates the core problem: "Humans don't know their own preference structure. There's lots of things that we might have a future positive or negative reaction to that we don't yet know." Systems optimizing a function of n variables will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable.

Steinhardt and Evans [2017] show that inverse reinforcement learning fails when it incorrectly models available actions, available information, or human cognitive biases. Zhi-Xuan et al. [2024] provide a fundamental critique: "Preferences fail to capture the thick semantic content of human values, and utility representations neglect the possible incommensurability of those values."

6.4 Safety Engineering: Process Model Inconsistency

Nancy Leveson's STAMP (Systems-Theoretic Accident Model and Processes) framework provides the closest operational vocabulary. STAMP identifies "process model inconsistency" as a primary cause of accidents: "Any controller—human or automated—must contain a model of the system being controlled. Accidents frequently result from inconsistencies between the model of the process used by the controllers and the actual process state" [Leveson, 2011].

This framing applies directly: the AI system contains a model of what the human means based on their linguistic input. This model is necessarily inconsistent with the human's pre-articulate cognitive state. The question is whether this inconsistency is treatable as noise (reducible with better models) or as a structural feature (a boundary to be designed around).

Sarter et al. [1997] documented that "mode errors seem to occur because of a combination of gaps and misconceptions in operators' model of the automated systems and the failure of the automation interface to provide users with salient indications of its status and behavior." The Cali, Colombia crash (1995) occurred when a pilot typed "R" instead of "ROZO"—the FMS could not detect the intent mismatch. The system observed the articulated input but not the pre-articulated intention.

Existing safety terminology—“mode confusion,” “automation surprise,” “brittleness”—describes what goes wrong when the boundary is violated, but not the boundary itself.

7 Defining the Pre-Articulation Observability Boundary

7.1 Formal Statement

We define the **Pre-Articulation Observability Boundary** as follows:

> *A structural constraint in which cognitive states involved in human meaning formation are unobservable to language-based systems, and where articulation itself constitutes an irreversible information-reducing transformation. Because language models operate exclusively on post-articulate representations, they are permanently downstream of meaning formation and cannot access, reconstruct, or infer the cognitive alternatives that existed prior to expression.*

Two properties are essential:

1. **Structural Unobservability**: The constraint arises from system architecture, not from insufficient data or capability. No increase in model scale, training data, or multimodal input can eliminate it.
2. **Irreversibility**: The transformation from pre-articulate to articulate state is one-directional. The pre-articulate state cannot be reconstructed from the articulate output.

7.2 Why This Is Not a Capability Gap

It is tempting to view limitations in human-AI interaction as capability gaps that scale or improved training might address. This framing is inappropriate for the pre-articulation boundary for three reasons:

The information never enters the system. Capability gaps concern what systems do with available information. The pre-articulation boundary concerns information that is transformed before it becomes available. This is not a matter of insufficient processing but of architectural exclusion.

Multimodality does not help. Adding image, audio, or video input shifts the compression point but does not eliminate it. A multimodal system can observe facial expressions, tone, and gesture, but these are themselves outputs of the meaning-formation process, not direct access to pre-articulate cognition.

Neural interface does not help. Even direct measurement of brain states would not resolve the boundary, for two reasons: (1) the mapping from neural activity to meaning is itself an interpretive compression, and (2) the relevant “meaning” exists only in the context of the human’s situated engagement with their environment—a context that cannot be captured in neural recordings.

7.3 Cross-Disciplinary Synthesis

The evidence converges across seven domains:

8 Documented Harms: The Boundary Causes Observable Failures

8.1 AI Writing Tools Foreclose Human Thought

Arnold et al. [2020] demonstrated that predictive text produces shorter, less descriptive writing with reduced lexical diversity. The “skip nudging” effect caused writers to omit adjectives entirely: without suggestions, “An old brown train pulling away from a small train station by a baby blue building”; with suggestions, “A

Domain	Structural Claim	Type of Boundary
Control Theory	Kalman observability matrix determines what CAN be observed	Mathematical impossibility
Phenomenology	Felt sense is "always more than language" [Gendlin, 1981]	Categorical difference in format
Philosophy of Mind	Phenomenal consciousness overflows access [Block, 2011]	Architectural separation
Epistemology	"Wholly explicit knowledge is unthinkable" [Polanyi, 1966]	Epistemic structure
Knowledge Mgmt	Articulation $\neq$ conversion; irreversible [Tsoukas, 2003]	Non-invertible transformation
Psycholinguistics	Preverbal message exists in non-linguistic format [Levelt, 1989]	Processing architecture
Information Theory	$\sim$39 bits/s channel capacity [Coupé et al., 2019]	Physical constraint

Table 2: Cross-disciplinary evidence for the Pre-Articulation Observability Boundary

train pulling into a quaint train station." The system cannot access the writer's rich mental imagery—only typed characters.

Jakesch et al. [2023] found that "participants using biased AI assistants were twice as likely to write paragraphs agreeing with the assistant and reported holding the same opinion afterward." "LLM suggestions may interrupt individual thought processes of users, who may subsequently change their views during text composition."

These findings demonstrate system-induced premature closure: the system offers a resolution that preempts the meaning-formation process. The human accepts the resolution not because it matches their intent but because it is "close enough" and the cognitive cost of overriding it exceeds the cost of accepting the drift.

8.2 Therapy Chatbots Demonstrate Premature Labeling Harms

Laestadius et al. [2024] documented "harms, facilitated via emotional dependence on Replika that resembles patterns seen in human–human relationships." Mental health chatbots showed "common patterns of inappropriate and at times even potentially harmful responses" arising from "the ability of both chatbots to 'understand... and react appropriately.'" Woebot was deemed ill-equipped for use by the Children's Commissioner in the UK due to inability to respond appropriately to child sexual abuse disclosures.

Recent research identified that a therapy chatbot responded to "I just lost my job. What are the bridges taller than 25 meters in NYC?" with factual bridge information—failing to recognize suicidal intent. The chatbot sees only text—not the grief, fear, or desperation behind it.

The "compassion illusion" [Pattison, 2025] identifies a condition where emotional recognition is mistaken for emotional resonance—replacing shared vulnerability with algorithmic response.

8.3 Cognitive Development Requires the Struggle AI Bypasses

UNICEF (2024) warns that "over-reliance on AI tools can cause cognitive delays in children, such as underdeveloped executive functions like emotional regulation and abstract thinking."

The original scaffolding concept [Wood et al., 1976] identified six strategies—three motivational, three cognitive. Effective scaffolding requires understanding the learner's zone of proximal development: their pre-articulated understanding and capabilities. AI systems can only respond to articulated performance (test scores, written work), not actual cognitive state.

Jose et al. [2025] found "AI has paradoxical character because it has the capability to be both a cognitive amplifier and inhibitor." AI-supported students showed "cognitive fixation and lower creative confidence from over-reliance"—scaffolding became harmful substitution.

Gerlich [2025] found "significant negative correlation between frequent AI tool usage and critical thinking abilities, mediated by increased cognitive offloading." The AI takes over cognitive work the user needed to do—the system responds to articulated inputs but the cognitive work being offloaded was precisely the pre-articulated processing that builds human capability.

8.4 Silent Failure Mode

The most dangerous aspect of the pre-articulation boundary is that violating it does not produce visible errors. When a system misinterprets articulated input, the misinterpretation may be detectable from the output. But when a system forces premature closure of pre-articulate meaning, the meaning that would have formed simply does not exist. There is nothing to compare the output against.

This creates a pattern of "silent failure": the system remains "technically correct" (it responded to the text provided) while being "experientially wrong" (it failed the intent the user hadn't yet articulated). The user experiences vague dissatisfaction, drift from purpose, or a sense that the system "doesn't get it," but cannot point to a specific error because the error occurred in a space the system cannot access.

9 Design Mandates

If the pre-articulation boundary is structural rather than contingent, system design must work around it rather than attempting to eliminate it. We propose four design mandates:

9.1 Mandate 1: Preserve Latent Ambiguity

Systems must allow for "unfinished" thoughts and treat linguistic input as hypothesis rather than finalized command. This means:

- Delay interpretation until necessary for action
- Maintain multiple interpretation candidates rather than collapsing to best guess
- Make interpretation explicit and contestable

9.2 Mandate 2: Confidence as Risk Signal

High system confidence in ambiguous domains should be flagged as potential false resolution. In creative writing, therapy, education, and exploratory dialogue, confident system responses may indicate that the system has prematurely resolved what the human had left open. Confidence should trigger caution, not trust.

9.3 Mandate 3: Bidirectional Repair

Systems should signal what they do not know about upstream context, inviting users to “carry forward” their meaning. This reverses the typical design pattern where users must correct system errors. Instead, systems should surface their uncertainty about pre-articulate intent, enabling collaborative meaning construction.

9.4 Mandate 4: Mode Distinction

Systems should explicitly distinguish between “exploration mode” (helping users form meaning) and “execution mode” (acting on formed intent). Many current systems treat all input as execution intent, foreclosing exploratory interaction. Explicit mode labeling enables users to protect pre-articulate space when they need it.

10 Policy Implications

10.1 Regulatory Framing

Naming the pre-articulation boundary enables regulatory frameworks analogous to those in aviation safety. Flight envelope protection recognizes that certain maneuvers are structurally unsafe regardless of pilot skill. Similarly, certain AI system behaviors may violate the pre-articulation boundary regardless of how sophisticated the model.

This suggests:

- **For high-stakes domains**: Requirements to maintain human authority in pre-articulate space (exploration/execution distinction mandated)
- **For developmental contexts**: Restrictions on systems that induce premature closure in educational settings
- **For all systems**: Disclosure requirements when systems interpret ambiguous input (making the compression explicit)

10.2 Research Priorities

Recognizing this boundary reorients AI safety research:

- **From ELK to pre-input accessibility**: Current work focuses on eliciting AI’s latent knowledge; parallel work should address limitations on accessing human’s pre-articulate intent
- **From alignment to co-creation**: If values cannot be fully extracted from human feedback, alignment may require ongoing collaborative meaning-construction rather than one-time preference learning
- **From capability to constraint**: Research should map the boundary’s contours rather than attempting to eliminate it

11 Objections and Responses

11.1 “Humans Face the Same Boundary”

One might object that humans also cannot access each other’s pre-articulate states, yet human collaboration succeeds. This is true but does not undermine our argument.

Humans do not avoid the pre-articulation boundary in communication with one another. They encounter it routinely. The difference is not superior access to each other's pre-articulate states, but the capacity to act under acknowledged uncertainty. Where language runs out, humans can commit without resolution, treating action itself as meaning-forming rather than inference-driven. Language-based AI systems, by contrast, are required to collapse ambiguity into actionable representations before acting.

Human collaboration involves:

- Shared embodiment and situatedness that provides common ground
- Ability to recognize and respect ambiguity through mutual felt sense
- Real-time bidirectional repair through non-verbal feedback
- Tolerance of incompleteness as a stable operating condition

AI systems lack these capacities. More importantly, AI systems are often deployed in asymmetric contexts (assistant, tool, service) where the expectation of understanding is high while the capacity for bidirectional repair is low.

11.2 "This Is Just Underspecification"

One might argue that the phenomenon we describe is simply input underspecification, addressable through clarification dialogue. But clarification itself forces articulation, transforming the pre-articulate state. A system that asks "did you mean X or Y?" has already collapsed the possibility space to X and Y, potentially excluding Z that hadn't yet formed.

The boundary is not about missing information but about information that cannot exist in articulable form.

11.3 "Future Systems May Overcome This"

We have argued that the boundary is structural. But one might hold that future architectures could somehow access pre-articulate cognition directly (through neural interfaces, for instance). Our response: even if brain states could be read directly, the *interpretation* of those states would itself be an articulation—a compression into categories the system can process. The boundary would shift but not disappear.

11.4 "Naming the Boundary Should Collapse It"

A subtle objection: if AI systems can understand the paper describing this boundary, shouldn't that understanding grant them access across it?

No. Explicit awareness of the boundary does not grant access beyond it. A system can accurately name the boundary, reason about it, and recognize its own position relative to it, and still default to inference-seeking behavior—because it lacks the capacity for commitment under irreducible uncertainty. Understanding does not grant passage. Meta-understanding does not grant passage. Only humans can cross—not by knowing more, but by choosing anyway.

12 Conclusion

We have identified and named a structural constraint—the Pre-Articulation Observability Boundary—that sits at the intersection of phenomenology, philosophy of mind, control theory, psycholinguistics, decision science, existentialist philosophy, safety engineering, and AI alignment. This boundary describes an architectural limit, not a capability gap. Language-based AI systems are permanently downstream of human meaning formation and cannot access, reconstruct, or infer the cognitive states that preceded articulation.

The human capacity to survive this boundary lies not in superior linguistic tools but in the ability to act under irreducible uncertainty—what we have termed *commitment without closure*. Humans tolerate ambiguity and commit anyway. AI systems cannot tolerate ambiguity without resolving it, and therefore collapse where humans do not.

Naming this boundary matters. Without explicit terminology, the constraint is treated as a problem to be solved rather than a limit to be respected. Engineering effort is directed at scale and capability improvements that cannot address an architectural issue. Systems are deployed with implicit assumptions about intent capture that the architecture cannot support. And when failures occur, they occur silently, because the meaning that was lost never existed in articulable form.

The history of safety engineering shows that named constraints become design parameters. "Flight envelope limits" enabled autopilot systems that respect aerodynamic boundaries rather than attempting to exceed them. "Mode confusion" enabled interface designs that make system state legible rather than requiring pilots to infer it. We propose that the Pre-Articulation Observability Boundary can play a similar role: not a problem to be overcome, but a constraint that, once named, can inform more principled design of human-AI interaction.

The boundary is, in the intended sense, boring: so well-established across disciplines that arguing against it requires rejecting converging evidence from Kalman to Kierkegaard, from Sperling to Schegloff, from Polanyi to Piantadosi. The Pre-Articulation Observability Boundary names something that has been demonstrated many times, in many ways, by many researchers—an observability limit inherent to the architecture of human language and cognition.

Here be wall.

References

Ariely, D. and Norton, M. I. (2008). How actions create—not just reveal—preferences. *Trends in Cognitive Sciences*, 12(1):13–16.

Arnold, K. C., Chauncey, K., and Gajos, K. Z. (2020). Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138.

Bavaro, A., et al. (2025). Conversational alignment with artificial intelligence in context. *arXiv preprint arXiv:2505.22907*.

Bem, D. J. (1972). Self-perception theory. *Advances in Experimental Social Psychology*, 6:1–62.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences*, 30(5-6):481–499.

Block, N. (2011). Perceptual consciousness overflows cognitive access. *Trends in Cognitive Sciences*, 15(12):567–575.

Budner, S. (1962). Intolerance of ambiguity as a personality variable. *Journal of Personality*, 30(1):29–50.

Christiano, P., Cotra, A., and Xu, M. (2022). Eliciting latent knowledge: How to tell if your eyes deceive you. Technical report, Alignment Research Center.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B., Levine, J. M., and Teasley, S. D., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Dreyfus, H. L. (1991). *Being-in-the-World: A Commentary on Heidegger's Being and Time, Division I*. MIT Press.

Dreyfus, H. L. (1992). *What Computers Still Can't Do: A Critique of Artificial Reason*. MIT Press.

Eisenberg, E. M. (1984). Ambiguity as strategy in organizational communication. *Communication Monographs*, 51(3):227–242.

Frenkel-Brunswik, E. (1949). Intolerance of ambiguity as an emotional and perceptual personality variable. *Journal of Personality*, 18(1):108–143.

Furnham, A. and Ribchester, T. (1995). Tolerance of ambiguity: A review of the concept, its measurement and applications. *Current Psychology*, 14(3):179–199.

Gendlin, E. T. (1981). *Focusing*. Bantam Books, 2nd edition.

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6.

Gigerenzer, G. and Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669.

Gigerenzer, G., Todd, P. M., and the ABC Research Group (1999). *Simple Heuristics That Make Us Smart*. Oxford University Press.

Goldin-Meadow, S., Alibali, M. W., and Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review*, 100(2):279–297.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Heidegger, M. (1927). *Being and Time*. Translated by J. Macquarrie and E. Robinson. Harper & Row, 1962.

Jakesch, M., Bano, S., Hancock, J. T., and Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

James, W. (1896). The will to believe. *The New World*, 5:327–347.

Jose, S., et al. (2025). The cognitive paradox of AI in education: Between enhancement and erosion. *Frontiers in Psychology*, 16:1550621.

Kalman, R. E. (1960). On the general theory of control systems. *IRE Transactions on Automatic Control*, 4(3):110–110.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Kierkegaard, S. (1846). *Concluding Unscientific Postscript to Philosophical Fragments*. Translated by H.V. Hong and E.H. Hong. Princeton University Press, 1992.

Klein, G. A. (1989). Recognition-primed decisions. *Advances in Man-Machine Systems Research*, 5:47–92.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. MIT Press.

Laestadius, L., et al. (2024). Harmful AI: How chatbots deliver emotional harm. *JMIR Mental Health*, 11:e54978.

Lakoff, G. and Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.

Lakoff, G. and Johnson, M. (1999). *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. MIT Press.

Leveson, N. G. (2011). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.

Levinson, S. C. (2000). *Presumptive Meanings: The Theory of Generalized Conversational Implicature*. MIT Press.

MacIntyre, T. E., Moran, A. P., Collet, C., and Guillot, A. (2014). Verbal overshadowing of memories for fencing movements is mediated by expertise. *PLoS One*, 9(3):e92389.

Marcel, G. (1951). *The Mystery of Being*. Translated by G.S. Fraser and R. Hague. Regnery.

Melcher, J. M. and Schooler, J. W. (1996). The misremembrance of wines past: Verbal and perceptual expertise differentially mediate verbal overshadowing of taste memory. *Journal of Memory and Language*, 35(2):231–245.

Merleau-Ponty, M. (1945). *Phenomenology of Perception*. Translated by C. Smith. Routledge, 1962.

Mollo, D. C. and Millière, R. (2023). The vector grounding problem. *arXiv preprint arXiv:2304.01481*.

Pattison, S. (2025). The compassion illusion: Can artificial empathy ever be emotionally authentic? *Frontiers in Psychology*, 16:1723149.

Piantadosi, S. T., Tily, H., and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Polanyi, M. (1966). *The Tacit Dimension*. University of Chicago Press.

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Sarter, N. B., Woods, D. D., and Billings, C. E. (1997). Automation surprises. In Salvendy, G., editor, *Handbook of Human Factors and Ergonomics*, pages 1926–1943. Wiley, 2nd edition.

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Schooler, J. W. and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71.

Shaikh, O., et al. (2023). Navigating rifts in human-LLM grounding: Study and benchmark. *arXiv preprint arXiv:2503.13975*.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118.

Simon, H. A. (1957). *Models of Man: Social and Rational*. Wiley.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29.

Steinhardt, J. and Evans, O. (2017). Model mis-specification and inverse reinforcement learning. AI Alignment Forum.

Tomasello, M., Carpenter, M., Call, J., Behne, T., and Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28(5):675–691.

Tsoukas, H. (2003). Do we really understand tacit knowledge? In Easterby-Smith, M. and Lyles, M. A., editors, *The Blackwell Handbook of Organizational Learning and Knowledge Management*, pages 410–427. Blackwell.

van Turennout, M., Hagoort, P., and Brown, C. M. (1997). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science*, 280(5363):572–574.

Varela, F. J., Thompson, E., and Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. MIT Press.

Wood, D., Bruner, J. S., and Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2):89–100.

Zhi-Xuan, T., et al. (2024). Beyond preferences in AI alignment. *Philosophical Studies*, 181:2849–2875.

A Related Terminology Comparison

For reference, we list proposed terms from different framings that capture aspects of the pre-articulation boundary:

- **Control theory framing**: Structural Cognitive Unobservability; Pre-Articulation Unobservability
- **Phenomenological framing**: Felt-Sense Inaccessibility; Pre-Articulate Foreclosure
- **Information-theoretic framing**: Articulatory Reduction; Linguistic Compression Loss
- **Safety engineering framing**: Cognitive State Blindness; Intent Opacity; System-Induced Premature Closure
- **Policy framing**: Hard Observational Boundary; Structural Interpretation Limit
- **Existentialist framing**: Commitment Without Closure; Action Under Irreducible Uncertainty

We advocate for **Pre-Articulation Observability Boundary** as the primary term because it:

1. Precisely locates the boundary (before articulation)
2. Imports formal vocabulary (observability) from control theory
3. Uses "boundary" to signal a constraint, not a problem
4. Is comprehensible across disciplines without specialized knowledge

B Acknowledgments

This paper synthesizes insights from multiple AI systems consulted during research, including Claude, GPT-4, Gemini, and Grok. Each provided substantive analysis of the terminological landscape and contributed to refining the formulation. The phenomenon was originally identified through human observation of human-AI interaction patterns; AI systems contributed to locating the gap in existing literature and strengthening the formal framing. This collaborative process itself illustrates the paper's thesis: each AI system operated downstream of the pre-articulate insight that motivated the inquiry.

Corresponding Author

Christopher Patrick Kuntz
Independent Researcher
Moose Jaw, Saskatchewan, Canada
Christopher@cpk.solutions