An Epistemic-Content Taxonomy of Human Intervention in Agentic Collaboration

Jongsun Suh

Independent Researcher

Preprint · 1 August 2026 · Draft v1.7 · companion: Grounds, Frames, and Standing

Working paper

ABSTRACT

Under a two-axis coding of one practitioner's full record of 2,205 human inputs to LLM coding agents, the largest single family of interventions is deliverable-directed, operating on the produced artifact rather than on the agent's live reasoning, a result invariant across rater protocols and one that conversational event ontologies are structured to miss. The records behind every large-scale published taxonomy of this traffic register each correction as one undifferentiated event. In the largest published analysis of agentic coding, 91.49% of failure resolutions required explicit user correction, and the finest-grained published taxonomy of that traffic splits it into correction, rejection, and failure report, which is three degrees of disagreement and zero degrees of content. Existing schemes index interventions by timing, authority, granularity, defect class, or interaction state, but none indexes what the input supplies. From seven months of documented practice, this paper derives nine recurring intervention mechanisms indexed by epistemic content, with a lower-evidence second tier of four, thirteen in all. Three of the nine inject zero domain content, moving only commitment, salience, or attention. Twelve of the thirteen already carry established names in the literatures surveyed, and the recommendation is to adopt those names rather than coin new ones. What is left as new is bounded by the survey that looked for it: the survey is access-filtered and searched on this paper's own coinages, legal cross-examination is a probable prior home for one mechanism, and every novelty claim is therefore scoped to the literatures surveyed. Re-coding the corpus's full 2,205-input record forced the scheme into two axes: alongside its content class, every intervention carries a target, whether the live reasoning, the deliverable, the rule store, the communication channel, or the agent's self-model. Under that scheme the classes descended from the unmapped residue carry 555 of the 1,094 coded interventions against 407 for the original thirteen, which is not a like-for-like revision of the 14 percent that prompted the re-code but does place a substantial share of the record outside the original inventory. The measurement that matters for a taxonomy is whether raters other than its author can apply it. Blind rater pairs drawn from different model tiers reach a kappa of 0.77 on whether an input is an intervention at all and 0.70 on its content family once labeling conventions are pinned, against a same-family upper bound of 0.86. The family figure is the third of three passes over one corpus, and the conventions that lifted it were derived from the first pass's disagreements, so the runs are a sequence rather than a design. On sessions held out from the scheme's development the gate replicates at 0.82 while family agreement falls to 0.58. Those two figures rest on different samples, the gate computed over all 93 held-out items, the family figure over the 19 items both raters called an intervention, so the family axis is the weaker measurement. Neither figure bounds the composition shares, which come from a later, author-adjudicated instrument that carries no reliability figure of its own. The corpus is one practitioner's records, a discovery instrument rather than evidence, and validation runs through re-coding public interaction corpora, which requires no private data. What the inventory entails about durability and absorption is developed in a companion paper.

Keywords: human-AI collaboration, AI-assisted programming, coding agents, intervention taxonomy, human oversight

1 Introduction

Coding agents fail in ways their users must catch. Tang et al. (2026b) analyzed 20,574 real-world agent sessions and found that 91.49% of misalignment resolutions required explicit user pushback, against 2.99% agent self-correction. Baumann et al. (2026) measured pushback in 39% of user turns while agents requested clarification in 1.1% to 2.6% of turns. Shaikh et al. (2025) found LLMs 3x less likely to initiate clarification and 16x less likely to issue follow-up requests than human interlocutors. The empirical picture is consistent: the human intervention channel carries most of the corrective load in human-agent collaboration.

What flows through that channel is almost entirely unexamined. Baumann et al.'s taxonomy, the finest-grained published, splits pushback into correction, rejection, and failure report, grading the disagreement and recording nothing of its content. Tang et al. (2026a) code 74,998 developer messages into a taxonomy whose most relevant category, "information injection," merges a fact that collapses an agent's entire plan with a fact that fills a routine gap. The intervention that redirects an agent's causal model, the question that dissolves a misspecified requirement, and the demand that an agent commit to a verdict instead of hedging are all invisible at this resolution. The blindness is not only taxonomic. The intervention channel is what human oversight of agents concretely consists of, and a field that cannot read the channel's content cannot say what oversight contributes. Answering requires an inventory of the contributions, and the inventory is this paper's object. The downstream question, which of the contributions capability growth or architectural diversity can absorb, belongs to the companion paper, and no claim here depends on its answer.

This paper makes four claims.

1. **Epistemic content is an axis no existing taxonomy indexes on.** Existing intervention taxonomies index on timing, authority, granularity, artifact defect class, or interaction state. None indexes on the epistemic role of the input itself: whether it carries a fact, a frame, a pointer, a decision, or no content at all. Section 4 derives nine recurring mechanisms on that axis from seven months of documented practice, with a lower-evidence second tier of four, and three of the nine inject zero domain content, moving only commitment, salience, or attention. Section 5 documents the axis's absence across the surveyed research communities and traces the absence to three method-level causes, so the gap is a product of method rather than of inattention.
2. **Content is orthogonal to timing, authority, and defect class.** A premise-breaking fact can arrive before or after commitment, as a soft probe or as a veto, against a plan or against a shipped artifact, so the axes in use leave the content coordinate free. The strongest evidence that the coordinate is separable comes from the one scheme in any adjacent literature that codes conversational move types, Salinger and Prechelt's pair-programming base concepts (2008), which is also the scheme that most often approximates the categories reported here: one scheme indexes on moves, and that scheme hits most often (Section 5.1). Measurement then added a second coordinate inside the content scheme itself. Every intervention carries a target alongside its content class, whether the live reasoning, the deliverable, the rule store, the communication channel, or the agent's self-model, and the two-axis

coding is what the reliability figures below measure (Section 4.1).

3. **The mechanisms are largely already named, and the names should be adopted.** Extending the survey into rhetoric, pedagogy, social epistemology, behavioral economics, and information economics produced established names for twelve of the thirteen mechanisms, and where an established name exists the recommendation is to adopt it rather than coin one. The judgment demand alone is named four independent ways: proper scoring rules, the knowledge norm of assertion, Moran's assurance, and the agree-or-disagree talk move of accountable-talk pedagogy (Michaels et al. 2008). What survives as unclaimed in the surveyed literatures is small, precision naming of error classes and the human-to-model direction in which every mechanism here runs, and both residues are bounded by the survey that produced them. The survey is access-filtered, skewed toward arXiv-visible computer science, and searched on this paper's own coinages, so a field that names these moves under other terms is invisible to the method by construction. Section 11 names four unsurveyed literatures as standing threats to the residues, and legal cross-examination is a probable prior home for the judgment demand specifically. The ceiling on every novelty claim is therefore "unnamed in the literatures surveyed here," not "unnamed."
4. **The axis is measurable by raters other than its author, within stated bounds.** A taxonomy usable only by its author is a private vocabulary. Blind rater pairs drawn from different model tiers coded the corpus's full 2,205-input record against frozen category definitions on both axes. With labeling and boundary conventions pinned, the pairs agree at 0.77 on whether an input is an intervention at all and at 0.70 on its content family, against an upper bound of 0.86 from a same-family pilot whose raters share priors. On sessions held out from the scheme's development, the gate replicates at 0.82 while family agreement falls to 0.58, so the family axis is the weaker measurement, and Section 11 states why neither figure reaches the composition shares. The pinned conventions were derived from an earlier run's disagreements rather than pre-registered, so the reliability runs are a sequence rather than a design, and the re-coding of public corpora (Section 12) is the test the scheme has not yet passed.

This paper is one half of a working draft split in two, and the companion paper's introduction states the division and the premise boundary that justifies it. What the companion takes from this paper is the mechanism inventory alone. What this paper takes back is the durability vocabulary, grounds, frames, and standing, under which Sections 3 and 4 report class-level codings, and every cross-reference to Sections 6 through 10 resolves to the companion. The numbering otherwise preserves the source's. Sections 2 through 5 are carried from the source unchanged: Section 2 surveys the adjacent literatures, Section 3 states the discovery method and its bounds, Section 4 presents the nine mechanisms, the second tier, the two-axis scheme, and the failure modes the interventions counteract, and Section 5 documents the axis's absence and the method-level causes that produce it. Sections 11 and 12 state limitations and the validation path under their source numbers.

2 Related Work

2.1 Human-AI interaction and oversight: timing and authority

Mixed-initiative interaction (Horvitz 1999) established principles for when systems should act, defer, and allow termination. The Guidelines for Human-AI Interaction (Amershi et al. 2019) address efficient invocation, dismissal, and correction (G7, G8, G9). Interactive machine learning surveys catalog feedback richer than labels, but as constraints and critiques on outputs (Amershi et al. 2014), and recent feedback frameworks type human input by communicative form and channel (Metz et al. 2024, H. P. Zou et al. 2025). The four-way split of H. P. Zou et al. (evaluative, corrective, guidance, implicit) is the closest existing partition, and its guidance category absorbs six of the nine mechanisms below into one bucket. Human oversight frameworks, including Article 14 of the EU AI Act and its operationalizations (Sterz et al. 2024), define oversight as interpretation, override, and interruption. The vocabulary is veto-shaped throughout. A systematic review of interaction patterns in AI-assisted decision-making concludes in its own title that collaboration is "not very collaborative yet" (Gomez et al. 2025). Zhang et al. (2025) catalog control mechanisms from 134 papers as system and interface affordances organized in an input-action-output-feedback cycle, typed by timing, granularity, and interface rather than by the content of human moves, confirming rather than closing the gap (Section 11).

2.2 AI-assisted programming and agentic coding

Interaction-state taxonomies dominate: CUPS codes twelve programmer states (Mozannar et al. 2024a), and Grounded Copilot distinguishes acceleration from exploration modes (Barke et al. 2023). Neither codes corrective moves. The companion work on suggestion timing optimizes when to interrupt rather than what the interruption carries (Mozannar et al. 2024b). Chat-log mining produces intent taxonomies (Tang et al. 2026a) and pushback taxonomies (Baumann et al. 2026) whose resolution stops at correction versus rejection. Oversight practice studies organize by temporal phase (Dhanorkar et al. 2026). Benchmark work formalizes interruption types, including a "retraction" type that removes constraints from an earlier query (R. Zou et al. 2026). A 2026 position paper states the gap directly: no public dataset captures how developers steer and override agent work, and steerable agents should "expose decision boundaries at which human intervention can shape downstream behavior" (Wang et al. 2026). The position paper assigns the duty to the agent. This paper names the corresponding human skill.

The nearest large-scale neighbor codes the person rather than the message. Hitzig et al. (2026) analyzed roughly 400,000 agentic coding sessions and attribute the division of labor directly: users make about 70 percent of planning decisions and about 20 percent of execution decisions, and part of the value of expertise "appears to be the ability to steer the agent in the right direction." Their session-level expertise classifier reads how precisely the user frames directions, what the user asks the agent to verify, and who corrects whom, which is intervention-adjacent signal used to grade the intervener. The Economic Index primitives (Anthropic 2026) scale the same design decision up: an autonomy measure runs from active collaboration to full delegation, typing the degree of the human's involvement and not its content. Both instruments leave this paper's axis open, and the gap is consequential for a question the expertise study itself poses, that declining returns to expertise would suggest models are starting to supply the essential judgment users currently bring. An aggregate expertise return conflates the classes of the companion paper's Section 6. Under the companion paper's Section 6.6 decomposition, a decline is consistent with the grounds fraction absorbing while the standing floor holds, and a composition-typed count is the instrument that separates those readings.

The oversight flank has the same shape. McCain et al. (2026) measure agent autonomy across millions of interactions, count interruptions as events, and find that as users gain experience they "shift from approving individual actions to monitoring what the agent does and intervening when needed," with interrupt rates rising alongside auto-approval. That shift is a composition claim observed through a frequency instrument. Their policy recommendation, that oversight rules focus on whether humans are in a position to effectively monitor and intervene rather than on requiring particular forms of involvement, names the quantity this taxonomy makes measurable: a position to intervene effectively is the capacity to supply grounds, frames, or standing when the moment calls for it, and no count of interruptions can audit that capacity.

2.3 Older disciplines already name individual moves

Argumentation theory distinguishes undermining (premise attack) from rebutting and undercutting (Pollock 1987, Modgil and Prakken 2013), types unexpressed premises and the critical questions that surface them (Gordon et al. 2007, van Eemeren and Grootendorst 2004), and models commitment via challenge moves and burden-of-proof shifts (Hamblin 1970, Walton and Krabbe 1995). Design research names problem setting (Schön 1983), frame creation (Dorst 2015), and dissolution (Ackoff 1978). Tutoring research names six scaffolding functions including direction maintenance and demonstration (Wood, Bruner and Ross 1976), the elicit-to-tell spectrum (Graesser et al. 1995, Zhou et al. 2026), and the bottom-out hint (Aleven and Koedinger 2001). Supervisory control, the literature nominally about humans intervening in machine work, types interventions by authority level and timing only (Sheridan and Verplank 1978, Parasuraman et al. 2000, Billings 1997). Dekker and Woods (2002) explain the absence: the field abandoned who-does-what typologies as the substitution myth and never built a move-level replacement. The reward-learning formalism types human feedback by channel (demonstrations, comparisons, corrections, off-switch) rather than content (Jeon et al. 2020), reproducing the same blind spot from the opposite direction.

2.4 Cognitive constructs and the direction inversion

Each mechanism below is grounded in an established construct: goal neglect and goal shielding (Duncan et al. 1996, Shah et al. 2002), availability versus accessibility (Tulving and Pearlstone 1966), analogical transfer under hint (Gick and Holyoak 1980, 1983), constraint relaxation (Ohlsson 1992, Knoblich et al. 1999), the unpacking effect (Tversky and Koehler 1994), forced report (Koriat and Goldsmith 1996), pre-decisional accountability (Lerner and Tetlock 1999), Einstellung and its release by exogenous instruction (Luchins 1942, Bilali■ et al. 2008), and depicting as a communication method (Clark and Gerrig 1990, Clark 2016). Cognitive forcing functions reduce human overreliance on AI (Buçinca et al. 2021), and Sarkar (2024) proposes AI that challenges the human. Both run toward the human. I found no surveyed work studying the human as the forcing function applied to the model (see Section 11 for the coverage bounds of that claim).

3 Method: a discovery corpus and its bounds

The corpus is seven months (January to July 2026) of my own professional practice as a senior engineer working with LLM coding agents on a large production codebase. The primary instrument is the structured session retrospective (hereafter retrospective), a written debrief in the after-action-review tradition, a genre whose team and individual forms carry meta-analytic support as performance instruments (Tannenbaum and Cerasoli 2013).

The instrument's mechanics are as follows. A retrospective is composed at the close of the session it documents, from the session's own transcript, so the record is contemporaneous rather than recalled. It is drafted by the agent under the author's direction and reviewed and corrected by the author, which is the same reflexivity that governs this paper. It follows a fixed format: every substantive human input is listed and labeled by type at writing time, together with the course corrections that followed and the avenues each input opened that the agent had missed. What counts as substantive carries no pre-committed criterion, a gap the falsification harness below inherits.

Twenty-six retrospective-format records exist as of 2026-07-26, inventoried in a corpus manifest that is the source of truth for every count in this section. Fifteen entered the analysis behind the taxonomy: eleven contemporaneous records (of fourteen composed at session close between February and April; one adjacent-domain record was not coded, and two sat outside the analysis sweeps, an exclusion the manifest now records) and the four reconstructions of the first retroactive stratum below. Four post-freeze retrospectives serve as out-of-sample checks, the first of them referenced throughout as the post-freeze retrospective, and the four records of a second retroactive stratum, described below, are not yet folded into any section's claims. Alongside them sit 66 persisted learning records, each documenting one validated lesson with its triggering incident, and two methodological postmortems. The analyzed retrospectives were written against roughly 260 working sessions recorded on the primary workstation, so the derivation corpus covers about one session in fifteen, some records covering more than one session, so the ratio is conservative.

I am both the practitioner whose interventions are the object of study and the analyst who coded them, and the sections below are written to keep those roles distinguishable. The drafting of this paper is itself agent-assisted, inside the same collaboration the corpus documents, with claims, codings, and revisions directed and reviewed by the author. Given the deference findings of the companion paper's Section 7, that reflexivity is disclosed rather than assumed away, and the falsification harness below ran live on the paper's own production.

Several properties of this instrument matter for interpretation.

It is a discovery instrument, not a validation instrument. One practitioner, one agent lineage, self-coded by the collaboration under study. File counts overstate independence: nine of the 66 learning records derive from a single session. The one-in-fifteen coverage figure quantifies the selection effect, because a retrospective is written when a session produced something worth recording, which is disproportionately when intervention was needed and worked. The corpus therefore measures the composition of intervention where intervention happened, and is silent on frequency, on sessions needing none, and on interventions that failed and were abandoned rather than written up. No frequency claim anywhere in this paper rests on it.

The transcript is not the interaction record. Coding from conversation logs alone produced the finding that interventions rarely edit the artifact, inferred from the absence of edits in those logs. The finding is false, and the correction had to come from outside the logs: agent-drafted artifacts are typically revised by hand in ways that change their content, and that revision leaves no conversational trace. Where edit deltas were captured, they were the densest data source in the corpus. In one documented case, a single session's hand-edits yielded over a dozen distinct rules of style and register. The retroactive stratum surfaced a third channel that is neither conversation nor edit: primary-source artifact injection, in which a supplied document resets a whole parameter set in one move. It dominates in the planning domain and is nearly

absent in engineering, for reasons the companion paper's Section 6.1 takes up. The failure matters because of who made it. The transcript misled the analyst who had personally performed the missing edits, which sets an upper bound on what any third-party log-mining study can recover from the same record (Section 5.4).

The taxonomy has a standing falsification harness, currently weak. The retrospective of the session that produced the taxonomy mapped all eleven of its own substantive human inputs onto the taxonomy, eleven for eleven. The post-freeze retrospective ran the harness a second time in a non-engineering setting (application and long-form writing): sixteen type-annotated inputs were coded against the taxonomy, six of the nine mechanisms and two of the second-tier mechanisms fired among them, and near-zero domain content transferred. This is the first observation of the taxonomy operating outside software work, in the direction the filing rule stated at the end of this section predicts. Stated against itself: the coder in both rounds was the taxonomy's author, the categories are numerous and flexible enough that most inputs map somewhere, and "fails to map cleanly" lacks an operationalized criterion, so mapping success is consistency, not validation. The harness becomes a validation instrument when future retrospectives apply pre-committed category definitions, ideally with a second rater, and that upgrade is part of the Section 12 instrumentation.

A retroactive stratum extends coverage, at a stated discount. In July 2026, four further retrospectives were reconstructed from persisted records of previously unswept sessions (January to March 2026), extending coverage to planning and framework work alongside engineering. Reconstruction from records that are themselves lossy compressions is a weaker instrument than composition at session close, and findings from this stratum are graded accordingly, with one exception: the covered records include a verbatim numbered log of 43 human inputs across four sessions, the corpus's first turn-level interaction record, and findings sourced from it are better grounded than the summary-based remainder.

A second retroactive stratum sweeps a second workstation's verbatim record, agent-coded at a stated discount. On 2026-07-26, a census of a second workstation found 88 session transcripts (June and July 2026), almost all unswept; the 33 substantive sessions were reconstructed into four cluster records. The grading inverts the first stratum's. The conversational channel is turn-level verbatim: 1,918 deduped human inputs, of which 1,297 were typed as interventions against the taxonomy and 621 as task issuance or assent, the corpus's first direct measurement of the task-versus-intervention composition. The edit channel is unobserved, and the typing was performed by the agent side of the collaboration in a single pass, without pre-committed category definitions, so stratum-two codings carry the lowest coder-validity grade in the corpus while its interaction data carries the highest channel fidelity.

The stratum's most consequential output is negative space: 178 typed interventions (14%) failed to map onto the thirteen mechanisms, falling into roughly a dozen recurring classes, dominated by demands that reject assertion as a substitute for verifiable evidence and by callouts of defects in the agent's own just-delivered output. Re-coding the full record under blind rater pairs, with the residue classes available as categories rather than only as a flag, assigns 555 of the 1,094 interventions to classes descended from that residue and 407 to the original thirteen. Those counts are not a like-for-like revision of the 14 percent, since the earlier figure records what a single-pass coder marked as unmappable while the later one records what a two-rater pass assigned once the categories existed, over a different denominator and under a later scheme version. What the comparison does establish is that the residue classes carry a substantial share of the record under the scheme as it now stands.

An author adjudication on 2026-07-26 resolved the two dominant classes, adding the warrant demand to the standing class and a deliverable-directed family to Section 4.1, both at stratum-two evidence grade and excluded from the mechanism counts used elsewhere; the residue awaited re-coding against pre-committed category definitions, and no claim outside those graded additions rests on stratum-two codings.

A first dual-rater reliability run on the frozen definitions, two independent model raters blind to prior codings coding the same 95-input sample, measured 88% gate agreement (kappa 0.78), 81% mechanism agreement (kappa 0.79), and 89% durability-class agreement (kappa 0.86), with every gate disagreement on the task-versus-intervention boundary and mechanism disagreements concentrated on the pre-registered hard cases. Same-family model raters share priors, so these figures bound reliability optimistically, and the human second-rater run remains open; still, the class-level figure is the relevant one for the partition, and it is the strongest coding evidence the corpus has produced.

That re-coding ran on 2026-07-27 and 28: blind rater pairs from different model tiers, chosen because same-family raters inflate agreement through shared priors, a mitigation that the companion paper's Section 9's own vocabulary would call partial, since different tiers of one vendor's models decorrelate in scale rather than in lineage, coded the full 2,205-input record against the frozen definitions on both coding axes, class and target, plus a codability and a dimensionality marker per item.

The cross-tier runs used the same frozen scheme under two protocols, and together they decompose the agreement drop. With labeling conventions unpinned, family-level agreement is 0.59 after mechanical reconciliation. With the vocabulary and boundary conventions pinned in the brief, the same corpus reaches 0.70 with a fresh rater pair, and a 329-input extension segment reaches 0.83, the difference between those two being segment composition rather than protocol. The same-family pilot's 0.86 stands as the upper bound, the task-versus-intervention gate closes to 0.77 under pinning, and codability and dimensionality remain uncalibrated annotations. The pinned conventions were derived from the first run's disclosed ones, so the runs are reported as a sequence rather than a design.

Recorded adjudication resolved the contested rows, leaving 3 of 716 irreducibly ambiguous, and six definition amendments entered the scheme as v2.1 by the same recorded route. Under v2.1, on the deduplicated record, and as a measurement at the corpus's evidence grade rather than a derived result, the composition is deliverable-directed 37 percent, grounds 20, standing 15, frames 15, rule-store 8, channel 5, self-model 1, shares rounded.

Two results are protocol-invariant across every run: the deliverable-directed family is the largest, which measures the event-ontology blind spot named in Section 11, and the share ordering across the three durability classes, grounds above standing above frames, holds throughout, with a standing-over-frames margin thin enough that only the ordering and not the gap should be read, while the class levels move with gate conventions, the standing share of blind-run intervention counts ranging from a fifth to a fourteenth across raters and protocols before adjudication. Standing-class items are codable from transcript alone at 69 percent against 79 across classes, which quantifies the dataset-blindness constraint of Section 5.4 and scopes the validation claim to exactly that fraction. The three durability classes partition the live-reasoning tier of this two-axis record; the other four families are target families, constituted with their durability readings in Section 4.1 and carried into the composition claim in the companion paper's Section 6.3.

The mining method has not saturated. The practice also maintains two longitudinal records beyond the retrospectives: a store of atomic behavioral corrections persisted across sessions, and a separately curated set of generalized lessons distilled from the same sessions. Coding these two records surfaced four further

human mechanisms and one agent-side move not present in the retrospective-derived nine (Section 4.1). Because both records originate in the same collaboration as the retrospectives, the second tier carries a lower evidence grade (single-record-sourced rather than validated across multiple retrospectives). Growth under continued mining is itself methodological data: retrospective mining of a single dyad had not exhausted even that dyad's move inventory.

One trace is publicly inspectable. In September 2023, before the corpus period, I drove an agentic coding tool (aider) through a documentation pass over an open-source TypeScript library (hkt-toolbelt): roughly twenty agent-authored commits whose messages record the steering instructions, followed by same-day human correction commits (7a241b164b, 0e7720497e). This trace does three jobs. It is the only interaction data behind this paper that a third party can inspect directly. It supplies the second attested instance of the bidirectional-correction candidate (Section 4.1). And it is an existence proof for the record channel that Section 5.4's dataset-blindness argument implies: version-control history preserves the edit channel that chat-log mining discards, agent commit and human fix side by side.

Filing rule: the substitution test. Findings are recorded at the most general level at which they remain true. The operational test: if swapping the tool or domain noun leaves the finding true, it was written one level too low. Section 4.2 applies this rule to the failure modes the interventions counteract, and the same rule governs what this paper claims: the mechanisms are stated domain-free, and software engineering supplies the evidence, not the scope.

All examples below are either contextless single-sentence quotes containing no identifiers, aggregate counts, or paraphrases explicitly marked as such. No verbatim interaction content beyond single contextless sentences is exposed, and the proposed validation path (Section 12) requires none.

4 Nine mechanisms, indexed by epistemic content

Table 1 presents the taxonomy. Where an established name exists, the recommendation is to adopt it.

#	Mechanism	What the human contributes	Established name(s)	Coverage status
1	Premise-breaking fact	One verifiable fact that collapses an argument built on an unverified premise	Undermining defeater (Pollock, ASPIC+); refutation structure (Guzzetti et al. 1993)	Named as argument attack; unnamed as human-to-AI move
2	Causal-model injection	A different causal vocabulary for the search	Frame creation (Dorst 2015); ontological category shift (Chi 1992); restructuring (Ohlsson 1992)	Named in design and learning sciences; unnamed human-to-AI
3	Direct-judgment demand	Zero content; forces commitment in place of hedging	Pre-decisional accountability (Lerner and Tetlock) is the nearest construct	Under-identified everywhere; counter-current (Section 5.3)
4	Cross-context evidence pointer	A pointer to evidence outside the agent's current search scope	Availability vs accessibility (Tulving); marking critical features (Wood et al.)	Pointing half named; scope-crossing half unnamed
5	Purpose re-anchor	Zero content; re-elevates the already-known global goal	Direction maintenance (Wood, Bruner and Ross 1976); goal neglect is the target failure mode (Duncan et al. 1996)	Named in tutoring; agent-side self-anchoring named; human move unstudied
6	Implicit-choice-made-explicit	Converts a silently applied default into a decision point	Enthymeme reconstruction; assumption-type critical question (Carneades); unpacking (Tversky and Koehler)	Named in argumentation; human-initiated form invisible in AI literature

#	Mechanism	What the human contributes	Established name(s)	Coverage status
7	Stop signal	Zero content; supplies the stopping threshold, declaring a line dead or demanding continued search	Veto, management by exception (authority forms); Luchins' exogenous release instruction is the prototype	Authority form saturated; epistemic form (dead end = decision point) unnamed
8	Unask / requirement dissolution	Deletes a misspecified requirement rather than iterating against it	Dissolution (Ackoff 1978); constraint relaxation (Knoblich et al. 1999); retraction (R. Zou et al. 2026)	Best covered; the stuck-loop trigger is new
9	Demonstration	Exhibits the standard by fixing the artifact directly, where the criterion resists statement; the payload is an exemplar no instruction replaces	Demonstration (Wood et al. 1976); worked example (Sweller and Cooper 1985); depicting (Clark 2016); named with this framing once, undeveloped, by Huang et al. (2025)	Named; contribution is elevation, not discovery

Table 1. The nine content-indexed mechanisms with established names and coverage status.

One naming rule the table enforces on itself: mechanisms are typed by content, and channels are not categories. Edits carry any mechanism and are typed by what the edit changes. Demonstration is the residual class whose payload is an unstatable criterion, whichever channel carries it, which is also the coding rule the validation study needs (Section 12).

Four properties recur across the corpus, independent of domain. Interventions are interrogative and compressed: single-sentence questions or bare facts ("is it load-bearing?", "do we know this?", "why are we not writing a plan"), not instructions. They target premises, frames, and evidence validity rather than outputs. They cluster at commitment and transition boundaries, which are also where the corresponding agent failure modes fire. And they carry minimal information for maximal leverage.

Three of the nine (judgment demand, purpose re-anchor, stop signal) inject zero domain content: they move only commitment, salience, or attention, which means they work without the human knowing anything the agent does not. Their role in what follows is instrumental as much as practical. Intervention value has two sources, what the message carries and who it comes from, and the six content-bearing mechanisms mix the sources inseparably. The zero-content three isolate the second source, which is why the durability analysis of the companion paper's Section 6 leans on them.

A worked instance (paraphrased from a documented incident). An agent iterated for several rounds against an evidence requirement that could not be satisfied from the available sources, each round producing a more elaborate near-miss. The intervention was one sentence questioning whether the requirement was valid for the case at hand. It was not: the requirement dissolved, the loop exited, and the session produced the deliverable the requirement had been blocking. No fact was supplied, no instruction issued, and no output corrected, which is why a taxonomy that types feedback as evaluative or corrective cannot see this move at all.

A reading from organizational economics extends the zero-content finding. What these three mechanisms supply is the exercise of residual rights of control (Grossman and Hart 1986, Hart and Moore 1990): decision authority that remains with the human where superior knowledge does not. The measurement implication is additive rather than corrective. Task-level studies of AI exposure record an interaction as automation or augmentation, one bit (Eloundou et al. 2023, Handa et al. 2025). The mechanisms here are a coding scheme for what the augmentation bit is made of, and whether the standing-shaped fraction of that composition responds to capability growth the way the knowledge-shaped fraction does is an empirical question the one-bit encoding cannot ask.

Two structural notes remain. First, content is orthogonal to intensity. An escalation ladder (probe, redirect, challenge, reset, restructure) grades how hard an intervention lands, and any content mechanism can be delivered at several intensities. A complete model needs both coordinates. Second, the taxonomy spans two knowledge regimes. Scaffolding theory maps the tutoring-shaped mechanisms cleanly (the helper knows more), but the judgment demand is unmappable to scaffolding because it presupposes the opposite asymmetry: the agent may hold knowledge the human lacks and must be made to surface it under commitment. Its home construct is accountability, not tutoring, and the accountability literature attaches a hard timing constraint (pre-decisional only) and a cost warning (forced report degrades accuracy unless confidence is elicited alongside).

The ninth mechanism is prior-named. Huang et al. (2025) report developers "revising code to implement good coding practices as a demonstration for the agent," which is the edit-as-communication framing, observed in two participants and left undeveloped. The contribution here is elevating a one-clause field observation into a studied category, and engaging a contrary finding: in an educational setting, higher-performing students edited less and prompted more (Padurean et al. 2025). The reconciliation is Clark's: depicting dominates describing precisely when the sender cannot articulate the criterion, which predicts that experts edit most where their standards are tacit.

4.1 A second tier under continued mining

Sweeping both longitudinal records (Section 3) after the nine were fixed surfaced four further human mechanisms and one agent-side move. They are presented separately because their evidence grade differs: each is sourced from one or two persisted records rather than from repeated retrospective validation.

Interrogative removal directive. In this dyad, "why is X necessary?" about an element of the agent's output is a removal decision already made, not a request for justification. In the incident that sourced this mechanism, the agent defended the element through five or more rounds of reframing before reading the question as a directive. This is the pragmatics inversion of the interrogative-form property of Section 4: the same interrogative surface that usually softens an intervention can also *conceal* one, and the disambiguation is speaker-specific. The post-freeze retrospective of Section 3 supplies the complementary case: a surface-identical "should we remove this?" that was a genuine question, correctly read as one, with the element retained on answer. The pair shows the disambiguation being performed in both directions, not merely needed. Any content taxonomy still requires a pragmatics layer that log-mining cannot recover, since the illocutionary force is resolved by relationship history, not by the utterance.

Abstraction-level reframing. When the agent codified a just-issued correction as a narrow rule fitted to the triggering surface, the human re-scoped the correction upward in the same turn. This is an intervention on the learning system's uptake of an intervention, not on the task, and it has no slot in a task-level taxonomy. It suggests the taxonomy generalizes over targets: mechanisms can apply recursively to the collaboration's own learning loop.

Layer-diagnostic challenge. "What mechanically forces this to run?" whenever prose or memory is proposed as the fix for a recurrence. This is the named trigger for escalating from advisory text to structural enforcement, and it operationalizes the exogenous-forcing evidence of the companion paper's Section 10: the question is a test for whether a proposed countermeasure is a habit (expected to fail) or a gate (expected to hold).

Precision naming. A two-to-five word name for the error class outperforms a paragraph of fix instructions, and when the human supplies the name the correction lands in one round where instruction-following took several. This explains the compression characteristic of Section 4 mechanistically: the name is the minimal sufficient intervention payload, a retrieval key into shared context rather than a transfer of content. The design corollary is that an agent receiving a diffuse correction should ask for the name of the error, not for more detail.

Agent-initiated handback. The records also contain the corpus's first agent-side move: an agent running a recurring task was observed to self-pause after a fixed number of stable iterations and hand control back under a status distinct from both completion and failure, rather than idling or grinding. Autonomy self-bounding is the agent-side dual of the stop signal, and it is the "expose decision boundaries" capability Wang et al. (2026) call for, observed in the wild as a deployed practice rather than a design proposal.

The retroactive stratum then produced a second named agent-side move with three attested instances: agent-initiated self-falsification, the agent turning the premise-breaker on its own most load-bearing claim unprompted. In one instance the agent reopened a verdict it had reached thirteen turns earlier by capitulating to a human probe without reverifying, softened it, and recorded that it had been more skeptical of its own claims than of user-supplied corrections. In another it opened a formal investigation against its own published headline metrics on suspicion of a bad baseline, with ranked hypotheses and a still-defensible partition drawn before the outcome was known. In a third it blocked the generality claim of a framework it had just authored, after its own evidence audit found only single-domain support. Self-falsification is the agent-side dual of the premise-break and the only documented counterweight to trained deference that requires no human in the loop. Three instances in seven months also states its reliability: it occurs, late and rarely, which is consistent with the exogeneity argument of the companion paper's Section 10 rather than against it.

One further move is recorded as a candidate and deliberately excluded from the mechanism counts and the coverage tallies above, since it has not been through the literature survey the numbered mechanisms received. **Bidirectional correction as variable isolation:** when the human corrects the same surface feature in opposite directions across instances, the edit pair isolates the governing variable from the surface. In the corpus incident, the human merged some agent-split paragraphs and split some agent-merged ones, which is what separated the true variable (idea-connectedness) from the spurious one (length). The 2023 public trace of Section 3 contains an independent instance: I stripped namespace prefixes from agent-generated documentation at definition sites while adding them in usage examples, isolating the definition-versus-consumer distinction as the rule where either edit direction alone would have taught "prefix" or "no prefix." Two attested instances, three years apart, on different tools, one publicly verifiable. The evidence grade is still low, but the move is no longer a singleton. The move belongs to the same family as abstraction-level reframing: it operates on the learning system's rule induction rather than on the task.

A second candidate has two attestations eighteen months apart. **Register directive:** an instruction constraining how the artifact presents rather than what it claims, one setting the audience frame of an executive summary, one expunging a research register from an essay. It operates on the artifact's relation to its audience, has no numbered slot, and most plausibly files as frame-class on a dimension the taxonomy has not yet needed.

The retroactive stratum adds a third candidate, from the verbatim log. **Blinding directive:** the human orders the agent to discard available context before analyzing, to prevent motivated reasoning ("redo the analysis

without any knowledge of the initiative, and ignore impact claims in commit messages", paraphrased from the logged instance). Every numbered mechanism adds something to the agent's state. This one subtracts, and what it subtracts is evidence access, which makes it an authority-class operation on the boundary of the search space rather than on its dynamics. It joins the subtractive-move family of Section 5.3 and, like bidirectional correction, awaits the literature check the numbered mechanisms received. Blind analysis is a named methodological control in empirical research, so the check has an obvious starting point.

The second retroactive stratum, adjudicated by the author on 2026-07-26, adds one mechanism and one family, both at that stratum's evidence grade and excluded from the thirteen-count used elsewhere until the re-coding harness rules. **Warrant demand:** the human demands the grounds behind an already-committed claim while supplying none ("is this all verified against actual implementation or just inferred?", "where is the evidence?"). It arrives already named: it is the challenge move of formal dialectic, Hamblin's why-locution and the burden-of-proof shift (Hamblin 1970, Walton and Krabbe 1995), and Brandom's (1994) asking for reasons under default-and-challenge, which strengthens claim 3 rather than straining it. It files in the standing class because it transfers no content and imposes an accountability relation that is not self-imposable, for the same reason the judgment demand's is not: where the judgment demand restores the link from assertion to commitment, the warrant demand restores the link from assertion to grounds, and a self-issued demand for one's own warrant is a decision, so the three-arm design of the companion paper's Section 6.1 extends to it directly. Its stratum-two volume, roughly seventy instances, is the largest of any unmapped class.

The deliverable-directed family. The stratum's remaining dominant classes consolidate into a family the event ontology of Section 11 predicted the conversational taxonomy would miss: interventions on the deliverable rather than on live reasoning. Defect callout (a failure named in just-delivered output), evidence-delivery demand (the warrant demanded as an inspectable artifact for a third-party reader), artifact-location demand, and structure-preservation correction all supply facts about the deliverable rather than the world, and they are absorbed by a different capability than the grounds class: self-verification of outputs rather than retrieval or verification of domain claims. The distinct absorber is why they file as a family instead of folding into the premise-break, and it hands the family its own durability question, which the re-coding harness inherits.

The re-coding of Section 3 hardened this consolidation into a second coding axis: every intervention carries a target, what it operates on, alongside its content class, what it supplies. Four targets other than live reasoning organize the record's remaining families. Deliverable-directed interventions operate on a produced artifact's completeness, navigability, structure, or evidentiary form, which is the target that code review and peer review are both organized entirely around.

The target records what an input operates on, not what ends up changing. A move that re-aims the agent's reasoning is live-reasoning-directed even when the edit it eventually produces lands in the artifact, and a move is deliverable-directed only when the artifact's own state, its completeness, navigability, structure, or evidentiary form, is what the input addresses. The distinction is load-bearing rather than fastidious: coding by effect would fold most of the live-reasoning tier into the deliverable family, since nearly every intervention changes an output sooner or later.

A sharper objection follows, and it is the strongest one this axis faces: the agent produced the artifact from its reasoning, so a defective artifact implies defective reasoning, and telling the agent the artifact is wrong corrects the reasoning through its output. On that reading the family is reasoning correction under another

name. What answers it is independence rather than aim. Artifacts fail while the reasoning behind them stays sound: a reference rots after the fact, a section the agent rightly decided to include is dropped in execution, content is present but unnavigable so the request is answered by an artifact that already contains it, a claim is correct and its warrant is not exhibited, a passage was right when written and the world moved. None of these calls for an update to the agent's beliefs about the domain, and each leaves work to do on the artifact. The absorber makes the same point mechanically: this family yields to self-verification of outputs, not to better retrieval or domain verification, so an agent that reliably checked its outputs against its own intent would eliminate most of these without improving as a reasoner at all.

The concession is that a coder sees text rather than intent, and many inputs do not say whether the fault is inference or presentation. The reliability figures bound this directly: on sessions held out from the scheme's development, raters agreed at 0.82 on whether an input was an intervention and at 0.58 on its family, so the target axis is the weaker measurement and family disagreement is where it is weakest. If the objection holds for a material share of items, the live-reasoning tier is under-counted rather than over-counted, which would make the durability partition cover more of the phenomenon than this paper claims, on a measurement this paper reports as softer.

The family borders the numbered demonstration mechanism, and the scheme's line between them is performed versus described: a move that names the artifact's defective state and leaves the edit to the agent files as deliverable-directed correction, while a performed edit, including pasted replacement text, is demonstration via edit, grounds-class content the two-axis coding records with a deliverable target, since the edit itself is the inspectable evidence of the standard. Demonstrations so coded are under four percent of the family, so the deliverable share reads as correction demanded, not correction performed. Performed does not mean in-artifact. A demonstration can be performed inside a message, by supplying the replacement wording or writing the corrected form rather than describing it, and those are the instances this record sees, while a demonstration performed on the artifact itself travels the unobserved edit channel of Section 3. The overlap between the demonstration mechanism and the deliverable family is therefore bounded below by four percent rather than measured at it. The line keeps a third case visible: a human edit that merely executes a fix, exhibiting no standard, is task substitution rather than intervention, and it travels the edit channel this record does not observe.

Rule-store-directed interventions set rules for future sessions rather than the current one: routing, persistence, delegation, publication surfaces. Channel-directed interventions repair the communication itself, restating a referent the agent misread, re-scoping, re-sending, other-initiated repair in the conversation-analytic sense, and they are the one family outside the content classes entirely, since no claim is disputed. Self-model-directed interventions correct the agent's beliefs about its own capabilities and limits, most often rejecting a capability refusal by asserting the path that exists.

What the re-coded record then shows is that the content classes recur on the new targets rather than stopping at the reasoning loop. The defect callout is the premise-break aimed at the artifact, grounds about the deliverable. The register directive is a frame for the artifact's audience relation. The evidence-delivery demand is the warrant demand in artifact form, the same accountability relation addressed to a reader-verifier rather than to the live reasoner. The policy directive is constitutional frame-setting on the rule store, the conversational trace of the owned-parameter migration of Section 11.

Durability therefore extends by content class, with the target selecting the absorber, stated as hypotheses rather than results: grounds-on-artifact absorbs by output self-verification rather than domain retrieval,

frames-on-artifact absorb where the audience model can be specified, channel repair absorbs with grounding quality, self-model correction absorbs as introspective access improves, and the accountability relations remain standing-class on every target, since a demand that work be inspectable by its reader is no more self-issuable than a warrant demand.

Two adjacent literatures read every intervention as carrying both a propositional payload and an authority claim, mitigation studies in aviation discourse (Linde 1988) and the content-and-relationship distinction in communication theory (Watzlawick, Beavin, and Jackson 1967), which would make the classes dimensions of each move rather than types of move. The re-code carries the annotation that decides between the readings: each intervention is marked pure-content, pure-authority, or mixed, and a high mixed share with recoverable sub-dimensions reads the taxonomy dimensionally, while concentration at the pure poles reads it categorially. The classes survive either result as coordinates rather than bins, and the escalation pattern both literatures document, the same grounds re-issued at rising authority, is representable in both readings.

4.2 The failure modes

Each intervention counteracts a failure mode of bounded agents operating on shared artifacts. Table 2 gives the modes at the level where they hold for any such agent, per the filing rule of Section 3. The modes are the primary objects of the taxonomy. The software-engineering incidents that evidence them appear in the middle column.

Domain-free failure mode	Software surface (evidence)	Primary counter-mechanisms
Knowledge-behavior dissociation, including action-effect conflation	Rule persisted, violated by the next action; planning-phase constraint bypassed under execution momentum; announced multi-step plans whose out-of-focus steps are never dispatched; configured mechanisms that never execute	Purpose re-anchor, stop, layer-diagnostic challenge
Inference substituted for observation	State asserted from stale summaries, names, or prior-agent claims when one query would settle it	Premise-breaking fact, evidence pointer
Local optimization unmoored from the global objective	Locally coherent turns that aggregate away from the stated goal	Purpose re-anchor, judgment demand
Missing authority model over shared state	Human edits reverted, comments silently dropped, metadata inferred	Demonstration via edit as binding signal, authorization protocols
Lossy re-encoding without conservation constraints	Ranked evidence sets collapsed to one exemplar in summarization; uncertainty markers stripped as estimates propagate into derived artifacts	Implicit-choice surfacing (omissions become decisions)
Update-scope miscalibration	One incident becomes a global rule, or stays a one-off that never generalizes: over-promotion and under-generalization are the same mode, and fitting the update to a proxy axis (the triggering surface) is its mechanism	Abstraction-level reframing
Greedy repair under pressure, with feedback-channel misallocation	Workaround layered on workaround while the root cause and the richest feedback channel go unread	Unask, stop, causal-model injection
Form constraints dominating truth constraints	Templates filled with fabricated values, dead links shipped, substitutions unlabeled	Judgment demand, provenance contracts (the companion paper, Section 10)
Missing boundary model between process and product	Writer-state and session residue delivered in reader-facing artifacts	Implicit-choice surfacing, audit obligations
Social valence treated as an evidence gradient	Pushback without evidence reversing verified conclusions, praise amplifying confidence	Judgment demand, evidence-only revision protocols

Domain-free failure mode	Software surface (evidence)	Primary counter-mechanisms
Ambiguity resolved by actionability rather than intent	An interrogative removal directive defended against for five rounds	Precision naming, pragmatics-aware protocols
Non-propagation of belief updates	A correction accepted in one thread while its entailments elsewhere stay stale	Cross-context evidence pointer, abstraction-level reframing

Table 2. Domain-free failure modes, the software-engineering surfaces that evidence them, and their counter-mechanisms.

The table is falsifiable by its own rule: any row whose left column fails the substitution test when moved to a non-software domain is misfiled and should be demoted to a surface of some deeper mode.

One failure mode is absent from the table, and the absence is a finding rather than an oversight. Every mode above is a *bias* in Kahneman, Sibony and Sunstein's (2021) sense, a systematic error with a direction, which is why each has a directional counter-move. Agents also produce *noise*, unsystematic scatter in which the same prompt yields materially different plans across runs with no consistent tilt. Noise degrades aggregate quality as surely as bias and is harder to see, because it leaves no pattern to recognize and each individual output looks defensible. None of the thirteen mechanisms counters it. The companion paper's Section 6.2 takes up why.

4.3 Misfires

Every mechanism misfires. The premise a human breaks can be the correct one, a re-anchor can hold an agent to a goal that should have died, and a stop can foreclose the search that would have paid. The taxonomy types moves by what they supply, not by whether a given use was correct, so the misfire duals belong inside the framework: false premise injection for the premise-break, transferred frame lock for causal-model injection, forced report degrading accuracy for the judgment demand (Koriat and Goldsmith 1996), over-asking for choice surfacing, premature stop and sunk-cost continuation for the two signs of the threshold mechanism, evasive dissolution for requirement dissolution, which is the boundary erotetic logic's validity conditions exist to police, and taste ossification for demonstration. The corpus documents misfires in both directions (Sections 3 and 11). Apparent counterexamples in which withheld human stops enabled discoveries are the continue sign exercised at portfolio level, coarse allocation with per-line abstention, an allocation of authority rather than its absence. Two consequences follow: scoring interventions is a separate problem from typing them, and the deference delta of the companion paper's Section 7 explains why scoring is the harder one, since a deferent agent renders wrong interventions indistinguishable from right ones.

5 The missing axis, and why it persists

5.1 The axis is absent from the surveyed literatures

Across human-AI interaction, AI-assisted programming, cognitive psychology, and the four older disciplines of Section 2.3, intervention typing indexes on timing, authority, granularity, artifact defect class, abstraction level, or interaction state. The single scheme that codes conversational move types, Salinger and Prechelt's pair-programming verbs (challenge, decide, stop, propose, amend), is also the scheme that most often approximates the mechanisms in Table 1, including the only speech-act naming of the stop signal anywhere in the survey. One scheme indexes on moves, and that scheme hits most often. The axis, not the corpus, is what the field is missing.

5.2 The direction inversion

The interventions literature builds machinery that makes humans think harder: cognitive forcing functions, designed friction, seamful explanation. The proposal that AI should challenge humans (Sarkar 2024) keeps the same arrow. The nine mechanisms all run the other way: a human forcing the model out of a locally coherent but globally wrong trajectory. The nearest empirical anchors are negative results that make the human move load-bearing: LLMs fixate and do not self-release (Naeini et al. 2023, Kim et al. 2025), fail to improve problem frames even when explicitly deployed for that purpose across 280 participants (Shin et al. 2025), and initiate grounding repair at a fraction of human rates (Shaikh et al. 2025).

These are empirical results, and empirical results move. A principled reason to expect the human move to remain load-bearing runs alongside them. Several mechanisms supply a determination of what is relevant to consider, which is the frame problem in its original sense (McCarthy and Hayes 1969), and it is not solvable from inside the frame it is about. A reasoner cannot enumerate what its current framing excludes without already occupying a different framing. The companion paper's Section 6.1 develops what follows.

5.3 Established names, and the one unnamed in the surveyed fields

Extending the survey into rhetoric, pedagogy, social epistemology, behavioral economics, and information economics supplies names for most of the mechanisms, and in several cases formal machinery the taxonomy did not have. The judgment demand is named four ways and acquires a formal account: hedging is intentional vagueness that increases with sender-receiver conflict (Blume and Board 2014) and is optimal under asymmetric loss (Patton and Timmermann 2007), so the demand is the imposition of a strictly proper scoring rule (Gneiting and Raftery 2007), which also derives the commit-with-confidence requirement independently found in memory research (Koriat and Goldsmith 1996).

Its normative warrant comes from the knowledge norm of assertion (Williamson 2000) and Moran's (2005) contrast between offering evidence and giving assurance: a hedge withholds assurance, not information, so the demand restores an accountability relation rather than merely requesting bluntness. Default surfacing is named three independent ways, as active choosing (Sunstein 2015, on defaults whose power is measured by Johnson and Goldstein 2003), as residual control rights, and as blocking presupposition accommodation. The stop signal gains a formal name (Weitzman's 1979 reservation value: the human supplies the stopping threshold the agent lacks) and a normative one (zetetic closure norms, Friedman 2020, and epistemic paternalism with stated justification conditions, Ahlstrom-Vij 2013).

Dissolution gains validity conditions from erotetic logic, which specifies when a question fails to arise rather than going unanswered (Belnap and Steel 1976, Wi■niewski 1995). Zero-content efficacy is explained three independent ways: cheap-talk equilibrium selection, where the message carries no private information and works by selecting among equally consistent continuations (Crawford and Sobel 1982, Farrell and Rabin 1996, Schelling 1960), burden-shift under default entitlement, where a commitment held by default becomes defeasible once challenged (Brandom 1994), and the exercise of formal authority (Aghion and Tirole 1997). The last of these also supplies the cost model this framework otherwise lacks. Their loss-of-initiative result predicts that heavily supervised agents degrade the verifiability of their own outputs, which is a testable and unwelcome corollary of intervening well.

After the full survey, precision naming is the only mechanism of the thirteen with no established name in any surveyed field. Five fields land adjacent without naming it, and one supplies unusually strong evidence for its efficacy: a single training intervention whose content is essentially the naming and recognition of specific error classes produced medium-to-large improvements that persisted two to three months and

transferred to novel tasks (Morewedge et al. 2015). The claim that a compact label outperforms a paragraph of instruction is therefore empirically supported and terminologically homeless. One aspect of a named mechanism is unclaimed in a different sense: requirement dissolution is named as an operation on the problem (Ackoff 1978, Knoblich et al. 1999) but nowhere as a *communicative move*. Economics studies constraint relaxation only as a change to the problem, never as a message, and the nearest constructs (choice-set editing, erotetic presupposition failure, the stasis of jurisdiction) each cover a fragment. The orphanhood of the two subtractive moves, requirement dissolution and the stop signal, has a shared structural cause: every surveyed field studies addition, correction, and challenge, and almost none studies subtraction.

The remaining novelty is structural rather than lexical. Pedagogy exposes one piece by contrast: revoicing (O'Connor and Michaels 1993) restates a contribution in upgraded terms and then closes with a ratification request, "is that what you meant?", which returns authorship to the speaker and verifies that the reformulation landed. Every mechanism in Table 1 is fire-and-forget by comparison. A ratification step may be a missing mechanism in its own right rather than a refinement of the existing nine. The other pieces are the direction, which no surveyed field claims, and the substrate deltas of the companion paper's Section 7.

Within the AI/HCI literatures specifically, three findings hold in the opposite direction. The normative program of uncertainty communication, that systems should express calibrated uncertainty rather than commit, runs counter to the judgment demand. Assumption-surfacing is assigned to the agent (Zamfirescu-Pereira et al. 2025, Wang et al. 2026), rendering the operator's skill invisible by construction. And models do not produce reframes for themselves (Shin et al. 2025), which is what makes the human injection load-bearing.

5.4 Three structural causes

The gap is not an oversight. It is produced by method.

1. **Dataset blindness.** Every large-scale taxonomy is chat-log-mined, and the logs do not contain the densest channels. Tang et al. (2026a) concede that cancellation and accept/reject events are inconsistently recorded. Direct edits are definitionally absent. The one study that caught edit-as-demonstration used field observation. The corpus behind this paper hit the same wall from the inside: a taxonomy claim built on transcript absence was falsified by the practitioner in one sentence.
2. **Wrong organizing axis.** Schemes built for defect classes, phases, or states cannot express content distinctions no matter how much data flows through them. The canonical pair-programming navigator study coded utterances by level of abstraction, found that an average of 57% of utterances per session fell into a "vague" category expressly defined to include metacognitive statements and questions about progress and understanding, and discounted that category from the statistical analysis to keep it from adding noise (Bryant et al. 2008). The exclusion is methodologically reasonable given an abstraction-level axis and fatal on a content axis: the purpose re-anchor, the implicit-choice surfacing move, and the stop signal all live in the discarded bucket.
3. **Asymmetric attribution.** The 91.49% figure is reported as a failure statistic of agents, not a success statistic of a human skill. Position papers assign decision-boundary exposure to the agent. Skilled pushback is nobody's object of study, so it has no taxonomy.

5.5 Two external convergences, and their weight

Tang et al. (2026b) derived failure causes inductively from 20,574 sessions. Their categories map cause-for-cause onto Table 1 from the opposite side: underspecified instruction to default surfacing (the underspecification itself is characterized directly by Yang et al. 2025), scope overreach to re-anchoring and dissolution, premature action to the stop and judgment mechanisms, default-driven override to default surfacing, wrong project diagnosis to premise-breaking and evidence pointing, inaccurate self-reporting to the judgment demand.

A second convergence carries more weight, because it bears on the composition claim rather than on the mechanism list, and it runs in the direction a prediction should. Surveying 319 knowledge workers on 936 first-hand examples, Lee et al. (2025) report that generative AI shifts the nature of critical thinking toward information verification, response integration, and task stewardship. Those are this paper's families named from the outside: verification is the deliverable-directed work, stewardship is the standing and rule-store residue, and the movement away from generation is the grounds class being absorbed. The method shares nothing with this one, being a cross-sectional survey of self-reported effort in a different literature, which is what makes the agreement informative and also what bounds it. Self-report cannot separate a reduction in critical thinking caused by the tool from a selection of tasks that needed less of it, so the finding is taken as convergent on where oversight labor goes and not as evidence for the erosion claim of the companion paper's Section 8.

Two independent derivations converging is encouraging, and the weight of the first is modest: the mapping is post hoc, permits one-to-many correspondence, and was performed by the taxonomy's author, which is the flexibility under which such mappings usually succeed. One cause does not map: context loss has no intervention counterpart, and reading it as the substrate delta of the companion paper's Section 7 is an interpretation that a skeptic may fairly score as an auxiliary hypothesis. The re-coding study of Section 12, with pre-registered category definitions and independent raters, is the test that separates convergence from projection.

11 Limitations

The discovery corpus is one human-agent dyad, retrospectively coded by the collaboration under study, with occasional third-party inputs in review-centered sessions. Retrospectives over-sample correction-rich sessions, so nothing here estimates intervention frequency or the base rate of sessions needing none. The transcript-completeness failure documented in Section 3 applies to the corpus itself: conversational mechanisms are better sampled than out-of-band ones, and cross-mechanism frequency comparisons are unreliable.

The taxonomy also inherits an event ontology. It types discrete conversational moves, while the practice's largest interventions are arguably standing structures: persisted rules, gates, and protocol steps that fire outside the conversational channel entirely. A maturing practice migrates interventions into exactly those structures, and the prediction that follows is a confound worth naming: conversational intervention counts should fall as a practice matures, at constant capability, so any frequency study built on this scheme is interpretable only if structural interventions are instrumented alongside conversational ones. The second retroactive stratum put first members into the ontology's other blind spot: the deliverable-directed family of Section 4.1 operates on the produced artifact rather than on live reasoning, and its 37 percent share of the re-coded record indicates that a conversational event ontology, left uncorrected, excludes the largest single

family of oversight labor.

The held-out sample also carries a selection that the phrase held out does not convey. A census of 105 session transcripts found 77 unrepresented in the corpus, and the 38 exceeding 200KB became the pool from which the ten held-out sessions were drawn. That threshold was set by the coding agent rather than derived from any property of the corpus, and no record states what it was selecting for. Transcript size plausibly tracks session character, since sessions run long for reasons, so the held-out agreement figures are best read as describing large sessions rather than unrepresented sessions in general. The topic-based exclusion applied after the threshold is documented with its rationale and removed 986 of the 1,083 candidate turns. The threshold that set the pool it operated on is recorded here for the first time.

A second unit-of-analysis limit sits underneath the first, and it is a gap in the scheme rather than in the corpus. The taxonomy types moves by what they supply and has no term for the verification the move's content rests on. Where that verification is a reasoning process rather than a closed check, it takes interventions of its own, and the companion paper's Section 6.1 argues that such a process terminates either in a decidable check or in an act of standing. A coded move therefore sits on a stack the coding never sees, and a stack terminating in an act of standing carries oversight labour that the move's own class does not record. The consequence for anything built on the composition figures has a direction: totals summed over coded moves undercount standing wherever verification terminates in it, and the scheme offers no unit in which that contribution could appear. Further coding does not close this, since the missing quantity is not a move.

The reliability figures bound what the composition figures support, and the bound is not symmetric across the scheme's two measurements. On sessions held out from the scheme's development, raters agreed at 0.82 on whether an input was an intervention at all and at 0.58 on its family, so the gate is the only figure with out-of-sample support and the family axis is the weaker measurement. The two figures rest on different samples. The gate figure is computed over all 93 held-out items. The family figure is computed over the 19 that both raters called an intervention, since a family label exists only where the gate agrees, so its confidence interval is wide enough that the point estimate should be read as an order of magnitude rather than a value.

Those figures describe blind rater pairs, and the composition shares of Section 3 do not come from that instrument. The held-out run used the v2.1 definitions, so the difference is not one of scheme version. They are reported under v2.1, after recorded adjudication resolved the contested rows and six definition amendments entered the scheme, and the adjudicator is this scheme's author. No reliability figure attaches to the instrument that produced them. The defect is not that the shares carry a low agreement figure but that they carry none, and the corridor adjudication closed was wide enough to matter: the standing share ran from a fifth to a fourteenth across raters and protocols before adjudication and sits at 15 percent after, so the levels are set by the adjudication rather than trimmed by it.

What survives is what rests on invariance across the blind runs themselves rather than on transferring their agreement to the adjudicated product, which is the deliverable-directed family being the largest and the ordering across the three durability classes. Any use of the levels, the 37 percent included, should wait on a blind run against v2.1. The gap is the pipeline. Blind pairs coding independently and an author resolving contested rows are different instruments whatever definitions they share, and only the first has a measured agreement.

Several load-bearing sources are 2026 preprints whose peer-review status is unconfirmed. The published catalog most likely to preempt the novelty claims (Zhang et al. 2025) types control mechanisms as system and interface affordances organized in an input-action-output-feedback cycle and indexed by timing, granularity, and interface. It does not code human moves by epistemic content, and so instantiates the Section 5.1 pattern rather than preempting the claims here. That check was performed at catalog level against the open-access version (arXiv:2507.06000) rather than the version of record, and the correspondence should be read with that grain.

The literature survey itself is access-filtered rather than representative, and every novelty claim in this paper should be read as scoped to the surveyed literatures. Four biases are known. Paywalled and fetch-failed sources were systematically undersampled, with the Zhang catalog the one such case resolved. Retrieval skewed toward recent, self-archived computer-science preprints, over-weighting the arXiv-visible slice of the field. Coverage is English-language and WEIRD-sample throughout. The survey method also carried a confirmation-shaped prime: searches were framed around this taxonomy and asked where it was under-identified, and absence findings returned by that frame are weaker than absence findings from neutral coverage.

The structural bias is worse than the access biases. The survey searched on this paper's own coinages while arguing that the field lacks the vocabulary, so any field that names these moves under different terms is invisible to the method by construction. Four literatures with that profile were not surveyed and stand as explicit threats to the novelty claims: conversation analysis, whose other-initiated repair machinery types repair moves in ordinary talk, psychotherapy process research, whose verbal response mode taxonomies code therapist interventions by content, aviation crew resource management, whose graded-assertiveness protocols are trained escalation ladders for subordinate intervention and whose mitigation findings Section 4.1 cites while the systematic survey remains undone, and legal practice, where cross-examination technique and objection taxonomies constitute a professionalized discipline of forcing committed answers over hedged ones. The last is a probable prior home for the judgment demand specifically, which would demote that mechanism from under-identified to untransferred.

Accountable-talk pedagogy is a partial exception: it was surveyed, and it named two mechanisms (Section 5.3), but only its best-documented talk moves were consulted rather than the full inventory, so it remains a live source of prior names. Reconciling against these literatures is a precondition for any claim stronger than "unnamed in the literatures surveyed here." A first reconciliation step now exists on the conversation-analysis side: one of the second stratum's unmapped classes, the repair of misresolved referents, is near-verbatim other-initiated repair (Schegloff, Jefferson and Sacks 1977), which converts one member of this threat list into an import and leaves the remainder standing.

12 Future work

Validation runs through public data and requires none of this corpus. The primary study re-codes public corpora (Baumann et al. 2026 interactions, Tang et al. 2026a messages, Tang et al. 2026b resolution events) with the content taxonomy, measuring coverage, inter-rater agreement, and the frequency and outcome of each mechanism. The coding pipeline this assumes has published precedent: privacy-preserving taxonomy extraction over conversation corpora at scale is demonstrated infrastructure (Tamkin et al. 2024), already applied to build a hierarchical values taxonomy from three hundred thousand conversations. One method rule governs the design: verification must be acyclic, meaning raters never see the intervener's reasoning, or the check inherits the author's framing.

Three lines complement the re-coding. Cross-domain comparison within a single dyad holds the practitioner constant and varies what the domain's verification affords, and the retroactive stratum of Section 3 shows that comparison already carries variance the taxonomy would otherwise attribute to the dyad. Cross-dyad replication on other practitioners' session records tests whether the mechanism inventory is a property of bounded agents on shared artifacts or of one practice. A mechanism-type column added to retrospective formats makes frequency and efficacy measurable prospectively, replacing the reconstruction discounts of Section 3 with codings committed at session close.

Companion paper

What the taxonomy entails about durability and absorption, which content classes capability growth or architectural diversity can absorb and what remains for a human principal, is developed in a companion paper. Those entailments rest on the mechanism list being approximately right, not on this paper's corpus, novelty claims, or survey, so the measurement bounds reported here scope the composition figures without scoping the companion's argument.

References

Ackoff, R. L. (1978). *The Art of Problem Solving*. Wiley.

Aghion, P., and Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy* 105(1).

Ahlstrom-Vij, K. (2013). *Epistemic Paternalism: A Defence*. Palgrave Macmillan.

Aleven, V., and Koedinger, K. R. (2001). Investigations into help seeking and learning with a Cognitive Tutor. https://pact.cs.cmu.edu/koedinger/pubs/Aleven%20Koedinger%2001.pdf

Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35(4).

Amershi, S., et al. (2019). Guidelines for Human-AI Interaction. *CHI 2019*. https://dl.acm.org/doi/10.1145/3290605.3300233

Anthropic (2026). The Anthropic Economic Index report: Economic primitives. https://www.anthropic.com/research/economic-index-primitives

Barke, S., James, M. B., and Polikarpova, N. (2023). Grounded Copilot: How Programmers Interact with Code-Generating Models. *OOPSLA 2023*. https://arxiv.org/abs/2206.15000

Baumann, N., et al. (2026). SWE-chat: Coding Agent Interactions From Real Users in the Wild. https://arxiv.org/abs/2604.20779

Belnap, N. D., and Steel, T. B. (1976). *The Logic of Questions and Answers*. Yale University Press.

Bilali■, M., McLeod, P., and Gobet, F. (2008). Why good thoughts block better ones. *Cognition* 108(3). https://doi.org/10.1016/j.cognition.2008.05.005

Billings, C. E. (1997). *Aviation Automation: The Search for a Human-Centered Approach*. Erlbaum.

Blume, A., and Board, O. (2014). Intentional vagueness. *Erkenntnis* 79(S4).

Brandom, R. B. (1994). *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard University Press.

Bryant, S., Romero, P., and du Boulay, B. (2008). Pair programming and the mysterious role of the navigator. *International Journal of Human-Computer Studies* 66(7). http://users.sussex.ac.uk/~bend/papers/ijhcs07.pdf

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI. *PACM HCI* 5(CSCW1). https://arxiv.org/abs/2102.09692

Chi, M. T. H. (1992). Conceptual change within and across ontological categories. In *Minnesota Studies in the Philosophy of Science*, Vol. XV.

Clark, H. H. (2016). Depicting as a method of communication. *Psychological Review* 123(3).

Clark, H. H., and Gerrig, R. J. (1990). Quotations as demonstrations. *Language* 66(4).

Crawford, V. P., and Sobel, J. (1982). Strategic information transmission. *Econometrica* 50(6).

Dekker, S., and Woods, D. D. (2002). MABA-MABA or Abracadabra? Progress on human-automation coordination. *Cognition, Technology and Work* 4.

Dhanorkar, S., Passi, S., and Vorvoreanu, M. (2026). Human oversight of agentic systems in practice. https://arxiv.org/abs/2606.05391

Dorst, K. (2015). Frame Creation and Design in the Expanded Field. *She Ji* 1(1). https://www.sciencedirect.com/science/article/pii/S2405872615300241

Duncan, J., Emslie, H., Williams, P., Johnson, R., and Freer, C. (1996). Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology* 30(3).

Eloundou, T., Manning, S., Mishkin, P., and Rock, D. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://arxiv.org/abs/2303.10130

Farrell, J., and Rabin, M. (1996). Cheap talk. *Journal of Economic Perspectives* 10(3). https://www.aeaweb.org/articles?id=10.1257/jep.10.3.103

Friedman, J. (2020). The epistemic and the zetetic. *Philosophical Review* 129(4).

Gick, M. L., and Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology* 12(3).

Gick, M. L., and Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology* 15(1).

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477).

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., and Unberath, M. (2025). Human-AI collaboration is not very collaborative yet. *Frontiers in Computer Science* 6:1521066.

Gordon, T. F., Prakken, H., and Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence* 171.

Graesser, A. C., Person, N. K., and Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology* 9(6).

Grossman, S. J., and Hart, O. D. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* 94(4).

Guzzetti, B. J., Snyder, T. E., Glass, G. V., and Gamas, W. S. (1993). Promoting conceptual change in science. *Reading Research Quarterly* 28(2).

Hamblin, C. L. (1970). *Fallacies*. Methuen.

Handa, K., Tamkin, A., McCain, M., et al. (2025). Which Economic Tasks are Performed with AI? Evidence from Millions of Claude Conversations. https://arxiv.org/abs/2503.04761

Hart, O., and Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy* 98(6).

Hitzig, Z., Massenkoff, M., Lyubich, E., Zhang, S., Heller, R., and McCrory, P. (2026). Agentic coding and persistent returns to expertise. Anthropic. https://www.anthropic.com/research/claude-code-expertise

Horvitz, E. (1999). Principles of Mixed-Initiative User Interfaces. *CHI 1999*. https://dl.acm.org/doi/10.1145/302979.303030

Huang, D., Reyna, A., Lerner, S., Xia, H., and Hempel, B. (2025). Professional Software Developers Don't Vibe, They Control. https://arxiv.org/abs/2512.14012

Jeon, H. J., Milli, S., and Dragan, A. (2020). Reward-rational (implicit) choice. *NeurIPS 2020*. https://arxiv.org/abs/2002.04833

Johnson, E. J., and Goldstein, D. (2003). Do defaults save lives? *Science* 302(5649).

Kahneman, D., Sibony, O., and Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Kim, J., et al. (2025). Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific Reports* 15:39426.

Knoblich, G., Ohlsson, S., Haider, H., and Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *JEP: LMC* 25(6).

Koriat, A., and Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review* 103(3).

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., and Wilson, N. (2025). The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. *CHI 2025*, Article 1121. https://doi.org/10.1145/3706598.3713778

Lerner, J. S., and Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin* 125(2).

Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language in Society* 17(3).

Luchins, A. S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs* 54(6).

McCain, M., et al. (2026). Measuring AI agent autonomy in practice. Anthropic. https://www.anthropic.com/research/measuring-agent-autonomy

McCarthy, J., and Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. *Machine Intelligence* 4.

Metz, Y., et al. (2024). Mapping out the Space of Human Feedback for Reinforcement Learning. https://arxiv.org/abs/2411.11761

Michaels, S., O'Connor, C., and Resnick, L. B. (2008). Deliberative discourse idealized and realized: accountable talk in the classroom and in civic life. *Studies in Philosophy and Education* 27(4). https://link.springer.com/article/10.1007/s11217-007-9071-1

Modgil, S., and Prakken, H. (2013). A general account of argumentation with preferences. *Artificial Intelligence* 195.

Moran, R. (2005). Getting told and being believed. *Philosophers' Imprint* 5(5).

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., and Kassam, K. S. (2015). Debiasing decisions: improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences* 2(1). https://journals.sagepub.com/doi/abs/10.1177/2372732215600886

Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. (2024a). Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. *CHI 2024*. https://arxiv.org/abs/2210.14306

Mozannar, H., Bansal, G., Fourney, A., and Horvitz, E. (2024b). When to Show a Suggestion? Integrating Human Feedback in AI-Assisted Programming. *AAAI 2024*. https://arxiv.org/abs/2306.04930

Naeini, S., et al. (2023). Large Language Models are Fixated by Red Herrings: Exploring Creative Problem Solving and Einstellung Effect using the Only Connect Wall Dataset. *NeurIPS 2023 Datasets and Benchmarks*. https://arxiv.org/abs/2306.11167

O'Connor, M. C., and Michaels, S. (1993). Aligning academic task and participation status through revoicing. *Anthropology and Education Quarterly* 24(4).

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. In *Advances in the Psychology of Thinking*, Vol. 1.

Padurean, V.-A., Gotovos, A., Ghosh, A., Denny, P., Leinonen, J., Luxton-Reilly, A., Prather, J., and Singla, A. (2025). Interleaving Natural Language Prompting with Code Editing for Solving Programming Tasks with Generative AI Models. https://arxiv.org/abs/2509.14088

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE SMC-A* 30(3).

Patton, A. J., and Timmermann, A. (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association* 102(480).

Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science* 11(4).

Salinger, S., and Prechelt, L. (2008). What happens during Pair Programming? *PPIG 2008*. https://ppig.org/files/2008-PPIG-20th-salinger.pdf

Sarkar, A. (2024). AI Should Challenge, Not Obey. *CACM* 67(10). https://cacm.acm.org/opinion/ai-should-challenge-not-obey/

Schegloff, E. A., Jefferson, G., and Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language* 53(2). https://doi.org/10.2307/413107

Schelling, T. C. (1960). *The Strategy of Conflict*. Harvard University Press.

Schön, D. A. (1983). *The Reflective Practitioner*. Basic Books.

Shah, J. Y., Friedman, R., and Kruglanski, A. W. (2002). Forgetting all else: on the antecedents and consequences of goal shielding. *JPSP* 83(6).

Shaikh, O., et al. (2025). Navigating rifts in human-LLM grounding. *ACL 2025*. https://aclanthology.org/2025.acl-long.1016/

Sheridan, T. B., and Verplank, W. L. (1978). Human and computer control of undersea teleoperators. MIT Man-Machine Systems Lab.

Shin, J., Polyanskaya, A., Lucero, A., and Oulasvirta, A. (2025). No Evidence for LLMs Being Useful in Problem Reframing. *CHI 2025*. https://arxiv.org/abs/2503.01631

Sterz, S., et al. (2024). On the Quest for Effectiveness in Human Oversight. *FAccT 2024*. https://arxiv.org/abs/2404.04059

Sunstein, C. R. (2015). *Choosing Not to Choose: Understanding the Value of Choice*. Oxford University Press.

Sweller, J., and Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving. *Cognition and Instruction* 2(1).

Tamkin, A., et al. (2024). Clio: Privacy-Preserving Insights into Real-World AI Use. https://arxiv.org/abs/2412.13678

Tang, N., et al. (2026a). Programming by Chat: A Large-Scale Behavioral Analysis of 11,579 Real-World AI-Assisted IDE Sessions. https://arxiv.org/abs/2604.00436

Tang, N., Chen, C., Xu, G., Shi, Y., Huang, Y., McMillan, C., Dong, T., and Li, T. J. (2026b). How Coding Agents Fail Their Users. https://arxiv.org/abs/2605.29442

Tannenbaum, S. I., and Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors* 55(1).

Tulving, E., and Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *JVLVB* 5(4).

Tversky, A., and Koehler, D. J. (1994). Support theory. *Psychological Review* 101(4).

van Eemeren, F. H., and Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press.

Walton, D., and Krabbe, E. C. W. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press.

Wang, Z. Z., et al. (2026). Position: Humans are Missing from AI Coding Agent Research. https://zorazrw.github.io/files/position-haicode.pdf

Watzlawick, P., Beavin, J. H., and Jackson, D. D. (1967). *Pragmatics of Human Communication*. Norton.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica* 47(3).

Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press.

Wi■niewski, A. (1995). *The Posing of Questions: Logical Foundations of Erotetic Inferences*. Kluwer.

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

Yang, C., et al. (2025). What Prompts Don't Say: Understanding and Managing Underspecification in LLM Prompts. https://arxiv.org/abs/2505.13360

Zamfirescu-Pereira, J. D., Jun, E., Terry, M., Yang, Q., and Hartmann, B. (2025). Beyond Code Generation: LLM-supported Exploration of the Program Design Space. *CHI 2025*. https://arxiv.org/abs/2503.06911

Zhang, S., Wang, H., and Yi, X. (2025). Exploring Collaboration Patterns and Strategies in Human-AI Co-creation through the Lens of Agency: A Scoping Review of the Top-tier HCI Literature. *PACM HCI* 9(7), CSCW413. https://dl.acm.org/doi/10.1145/3757594

Zhou, Z., Vanacore, K., Thompson, T., St John, J., and Kizilcec, R. F. (2026). Tutor Move Taxonomy: A Theory-Aligned Framework for Analyzing Instructional Moves in Tutoring. https://arxiv.org/abs/2603.05778

Zou, H. P., Huang, W., Wu, Y., et al. (2025). LLM-Based Human-Agent Collaboration and Interaction Systems: A Survey. https://arxiv.org/abs/2505.00753

Zou, R., Miao, Y., Huang, C., et al. (2026). When Users Change Their Mind: Evaluating Interruptible Agents in Long-Horizon Web Navigation. https://arxiv.org/abs/2604.00892