Grounds, Frames, and Standing: Human Interventions in Agentic Work and the Limits of Delegation

Jongsun Suh

Independent Researcher

Preprint · 26 July 2026 · Draft v1.1

Position paper with embedded case study

ABSTRACT

When, if ever, is human intervention required for an AI agent to operate successfully? The question is empirical, and the field cannot answer it, because its record of intervention is content-blind: in agentic coding, the best-measured setting, 91.49% of failure resolutions in the largest published analysis required explicit user correction, against 2.99% agent self-correction, yet every record registers the correction as a single undifferentiated event. Frequency says nothing about which interventions capability growth can remove. Existing taxonomies index interventions by timing, authority level, granularity, defect class, or interaction state. The missing axis is epistemic content: whether the input supplies a fact, a frame, a pointer, a decision, or nothing but a demand to commit. From seven months of my own practice with LLM coding agents, documented in structured session retrospectives, I derive nine recurring mechanisms indexed by that content, with a lower-evidence second tier. Three of the nine carry no domain content, moving only commitment, salience, or attention: authority exercised, not knowledge transferred. What a mechanism supplies determines what can absorb it. Mechanisms supplying grounds (evidence bearing on entitlement to conclusions) are absorbable by capability growth. Mechanisms supplying frames (replacements for the hypothesis space itself) yield only to a second reasoner, which need not be human. Mechanisms supplying standing require a principal, and standing cannot be self-conferred, so no capability increment absorbs them. Stated formally under incomplete contracting, expected loss from a misset delegation threshold converges to a strictly positive floor equal to the conditional dispersion of the principal's threshold, the compositional prediction follows as a corollary, and the third class is exposed as conditional on non-enumerability that persists under learning. The partition is an oversight result: a stop internalized as a preference transmits no standing, the delegation-side form of the corrigibility problem, and trained deference that rewards wrong corrections degrades the oversight signal itself. In the limit, an agent granted authority, accountability, and diverse priors absorbs everything but proprietorship: administering a success criterion is delegable, originating it is not, and the residue rests with whoever bears the outcomes. One institutional corollary: every external lever of control, liability included, presupposes a party that can be worse off, which derives the current priority of training-side alignment from institutional structure rather than capability risk. Across eleven literatures, twelve of thirteen mechanisms carry established names, and the taxonomy works as a concordance across fields that do not cite one another. What no surveyed literature claims is the direction: every adjacent field builds machinery that makes humans think harder, and none studies the human as the forcing function applied to a machine's live reasoning. The corpus is one practitioner's records, a discovery instrument rather than evidence. Validation is specified as re-coding public interaction corpora, where cause-for-cause correspondence with an independently derived 20,574-session failure taxonomy already provides preliminary convergence.

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: three degrees of disagreement, zero degrees of 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, which parts of it could be absorbed by capability or architecture, or what remains.

This paper makes six claims.

1. **Epistemic content is the missing axis.** 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. The one scheme in any adjacent literature that codes conversational move types, Salinger and Prechelt's pair-programming base concepts (2008), is also the scheme that most often approximates the categories reported here. That correlation is the argument for the axis.
2. **The direction is unoccupied.** Cognitive forcing functions (Buçinca et al. 2021), designed friction, and provocateur-AI proposals (Sarkar 2024) all run from system design toward the human. A taxonomy of human moves that force the model to reason better inverts that arrow, and no surveyed work occupies the inverted direction.
3. **The mechanisms are old. The assembly and the direction are not.** Extending the survey into rhetoric, pedagogy, social epistemology, behavioral economics, and information economics produced established names for twelve of thirteen mechanisms. 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), and hedging turns out to be formally characterized as optimal under an asymmetric loss function, making the demand the imposition of a proper one. What survives as unclaimed: precision naming of error classes (adjacent constructs in five fields, named in none), the injection of a replacement causal model into another reasoner mid-process, and above all the direction. Every adjacent literature runs

hearer-evaluating-speaker or architect-nudging-human, and none treats the human correcting a machine's live reasoning as its object.

4. **The substrate changes the dosing, not the moves.** Seven of nine mechanisms have exact cognitive-psychology analogs discovered on humans, so the moves are properties of bounded intelligence generally. What is specific to current LLM systems is a set of five deltas (non-fadeable scaffolding, bimodal retention, co-occurring opposed anchoring dispositions, a budget-based stopping rationale, and training-installed deference) plus a class of inverted-sign constructs that make naive imports from human-factors research actively misleading.
5. **The mechanisms sort by durability, and the sorting is derivable rather than observed.** What a mechanism supplies determines whether capability growth can absorb it, and the mechanisms supply three different kinds of thing: grounds, frames, or standing. Grounds-supplying moves are contingently human, because grounds live in the world and better verification reaches them. Frame-supplying moves require a reasoner outside the current frame, so they require an other but not necessarily a human. The three zero-content moves supply standing rather than content, and standing cannot be self-conferred, so no capability increment and no architecture absorbs them. This partition follows from the mechanism list plus established theory, holds independently of the corpus, and predicts that aggregate measures of AI exposure will report stability across periods in which the composition of human work inverts (Section 6). Section 6.6 states it as a decomposition under incomplete contracting, where expected loss from a misset threshold converges to a strictly positive floor given by the conditional dispersion of the principal's threshold, the composition claim falls out as a corollary, and the third class is exposed as a conditional result whose antecedent is non-enumerability that persists under learning. Its oversight consequences, from the delegation-side form of the corrigibility problem to the corruption of the oversight signal by model deference, are drawn in Section 9, and its limit case reduces to a custody and proprietorship split developed in Section 6.5.
6. **Validation must come from public data.** The discovery corpus is one practitioner's retrospectives: n=1, self-coded, selection-biased, covering roughly one working session in fifteen. The proposed validation study re-codes existing public corpora (Baumann et al. 2026, Tang et al. 2026a, 2026b) with the content taxonomy. The convergence already visible between this taxonomy and Tang et al.'s independently derived failure causes suggests the exercise will be productive.

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.

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 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 edited significantly by hand, and the hand-editing 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 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 real 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. Its 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. An author adjudication on 2026-07-26 resolved the two dominant classes, adding the warrant demand to the standing class and an output-directed family to Section 4.1, both at stratum-two evidence grade and excluded from the mechanism counts used elsewhere; the residue awaits 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.

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
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

#	Mechanism	What the human contributes	Established name(s)	Coverage status
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 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 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 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 Section 6.1 extends to it directly. Its stratum-two volume, roughly seventy instances, is the largest of any unmapped class.

The output-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 work-product 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.

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

Domain-free failure mode	Software surface (evidence)	Primary counter-mechanisms
Form constraints dominating truth constraints	Templates filled with fabricated values, dead links shipped, substitutions unlabeled	Judgment demand, provenance contracts (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
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. 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 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 Evidence of absence

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. Section 6.1 develops what follows.

5.3 Established names, and the one that remains unnamed

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 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 Convergent validity

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. Two independent derivations converging is encouraging, and the weight 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 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.

6 What the taxonomy entails

The taxonomy was derived from a corpus that cannot support generalization. Its entailments do not inherit that limitation. This section states results that follow from the mechanism list together with established theory, and that stand or fall independently of how representative the corpus is. Their premise is only that the mechanism list is approximately right, for which the convergence of Section 5.5 with an independently derived 20,574-session failure taxonomy is the current evidence. Whoever assembled the list, and in whatever domain, the following hold.

6.1 Three durability classes

Mechanisms differ in what they supply, and what they supply determines whether capability growth can absorb them.

Grounds-supplying mechanisms (the premise-breaking fact, the cross-context evidence pointer, demonstration via direct edit) supply grounds: considerations that revoke or confer entitlement to conclusions. Calling this class informational would mistake the payload for the action. A premise-breaking fact is an undermining defeater, and its work is revocation: one small verifiable fact withdraws entitlement from every conclusion resting on the defeated premise, so the effect is structural while the content is incidental. The evidence pointer supplies no content at all, only the location of grounds outside the agent's search scope, and the demonstration exhibits a governing standard as inspectable evidence rather than describing it. What unites the class is that its grounds live in the world. They are reachable in principle by

better retrieval, wider context, and stronger verification, which is why this class, alone among the three, is only contingently human. The contingency is domain-relative, and the corpus measures it directly. In engineering, where compilers, tests, and live queries put ground truth within the agent's reach, most grounds supply is verification the agent could have run. In the corpus's planning domain, ground truth lives in documents only the human holds, the dominant intervention becomes primary-source artifact injection, and the grounds class has no self-servable component at all. Absorbability of the grounds class is therefore a function of the domain's verifier availability, not of capability alone.

Frame-supplying mechanisms (causal-model injection, dissolution, implicit-choice surfacing) do something stronger than mark relevance. They replace the space within which relevance is evaluated. Causal-model injection swaps the vocabulary the search runs in, dissolution deletes a requirement constitutive of the problem statement, and implicit-choice surfacing converts a fixed piece of background into a live decision. Ranking hypotheses inside a space is ordinary inference, and a stronger reasoner does it better. Replacing the space is what Section 5.2 argues no reasoner can do for itself, since enumerating what a framing excludes already requires standing outside it. The operation is the classical second level of search: not search within a problem space but search of the space of problem spaces (Newell and Simon 1972, Kaplan and Simon 1990). This class is therefore not absorbable by a single reasoner becoming more capable, but it is absorbable by architecture, since a second reasoner with genuinely different priors performs the same function. What it requires is an other, not specifically a human.

Standing-supplying mechanisms are the three zero-content ones (judgment demand, purpose re-anchor, stop signal), and they differ in kind. They transfer no domain content, which raises an obvious question. If the message carries nothing the agent did not already have, why can the agent not issue it to itself? Because what transfers is not content but standing. On the property-rights reading adopted in Section 4, these are exercises of residual control rights (Grossman and Hart 1986, Hart and Moore 1990), and residual rights are defined as those remaining with the principal because they cannot be contracted away. A stop signal an agent issues to itself is not a stop signal but a decision. The entire content of the stop is that it originates with the party holding standing to end the inquiry. The judgment demand likewise imposes an accountability relation that is not self-imposable, because a promise extracted from oneself is not assurance in Moran's (2005) sense. There is no second party to whom one becomes answerable. The warrant demand of Section 4.1, this class's stratum-two addition, runs on the same asymmetry, restoring the assertion-to-grounds link where the judgment demand restores assertion-to-commitment. The purpose re-anchor re-asserts whose purpose governs, which presupposes a purpose-holder distinct from the executor. The claim is not circular, although it can look definitional. What the stop transfers is a threshold set by the principal's priorities across everything outside the task, which the executor can estimate but cannot authoritatively set, in the same sense that a reservation value belongs to the searcher and not to the search (Weitzman 1979). The supply is signed: the same authority that declares a line dead can demand search past an answer the agent was prepared to accept, and the corpus records both directions. Stop and continue are one mechanism, the setting of a threshold the executor does not own. The employment relation was formalized on exactly this object: Simon (1951) models employment as the purchase of authority over a zone of acceptance rather than of specified actions, and the three zero-content mechanisms are exercises within that zone, observed at message granularity. Capability growth therefore does not absorb this class, and architecture does not absorb it either, in a precise sense that distinguishes it from the frame class. The frame class needs only a peer, and any second reasoner supplies one. The standing class needs an asymmetry, and architecture can relocate the principal role to a different holder but cannot eliminate the role. The class dissolves only if the principal relation itself dissolves, which is a governance question and

not a technical one.

Table 3 summarizes the partition.

Class	Mechanisms	What transfers	Absorbed by
Grounds-supplying	Premise-breaking fact; cross-context evidence pointer; demonstration via edit	Grounds: considerations that revoke or confer entitlement to conclusions	Capability growth: retrieval, context, verification
Frame-supplying	Causal-model injection; unask / dissolution; implicit-choice surfacing	A replacement for the space of hypotheses itself	Architecture: any second reasoner with genuinely different priors
Standing-supplying	Judgment demand; purpose re-anchor; stop signal	Standing, not content	Neither: requires a principal, a role architecture can relocate but not eliminate

Table 3. The durability partition: what each class transfers and what can absorb it.

The partition is a prediction with a direct falsifier. A system that reliably self-generates well-timed stops, judgment commitments, and purpose re-anchors with no external principal would show the third class misidentified. The falsifier also adjudicates among readings of why zero-content moves work, and the candidate readings are game theory's. The message may be bare timing information, learnable from the principal's intervention history, in which case an agent trained on that history comes to issue equivalent moves to itself. It may be a correlation device in the sunspot sense (Cass and Shell 1983), in which case any public signal coordinates equally well and the sender is irrelevant. Or its force may live in the delegation relation itself. A three-arm design separates them: identical contentless interrupts issued by the agent to itself, by a script with no principal behind it, and by the principal. Success in the self arm collapses the standing class into the grounds class. Success in the script arm shows that delivering the interrupt mechanizes while leaving threshold ownership untouched, since a script that says stop is enforcing a threshold someone set. Only principal-arm superiority preserves the strong authority reading, and the partition predicts it. What the prediction does not assert matters equally. It does not say a *human* is permanently required. It says a *principal* is. Any arrangement conferring standing satisfies it, whether a delegating agent, an institutional rule, or a human operator, which relocates the question of durable human contribution from capability forecasting to the design of authority relations. Orchestration settings pose the mediated form of that question: when a human delegates to an agent that supervises other agents, standing is chained, and whether the standing mechanisms survive transmission through a delegate is unstudied.

6.2 A channel constraint, not merely a gap

That no mechanism counters noise (Section 4.2) is a constraint on the channel, not a gap in the list. Conversational interventions are directional. Each carries a target and a direction of correction, which is what makes the content axis meaningful in the first place. Directional instruments correct directional error, and unsystematic scatter has no direction to oppose, so it is unreachable by the whole channel rather than merely unaddressed by these thirteen instances. Adding mechanisms will not close it. The known remedies for noise in human judgment are structural rather than conversational, decomposing a judgment into independently scored factors and delaying integration until the components are complete (Dawes 1979, Einhorn 1972). A complete account of human contribution therefore requires a second and disjoint family of moves, operating on the process that generates outputs rather than on any output, and the two families are complements addressing orthogonal error classes. That division predicts, and may explain, the mutual silence between the intervention literature surveyed here and the evaluation and scaffolding literatures.

6.3 The augmentation bit has a measurable composition

Task-level studies of AI exposure record an interaction as automation or augmentation: one bit, the encoding used by exposure indices built on the task-based framework of labor economics (Autor 2015, Acemoglu and Restrepo 2019, Eloundou et al. 2023) and by direct measurement on live assistant traffic (Handa et al. 2025). Section 4 argued that the taxonomy supplies a coding scheme for what that bit is made of. Section 6.1 turns this into a testable claim about how the composition moves. The grounds-shaped fraction of human input should decline as capability grows. The standing-shaped fraction should not, because no capability increment confers standing. Aggregate exposure measures cannot detect this, since the bit is unchanged while its composition inverts, which means task-level AI-exposure studies may report stability across a period in which the character of human work changed substantially. The measurement is available on public interaction corpora and requires no private data, and it is the strongest reason to prefer a content taxonomy over a finer-grained intensity scale.

6.4 The operator reading

Every mechanism in the taxonomy can be rewritten as an operator on the agent's search state: the candidate space it is currently working, the boundary of what is reachable from its current context, the generator (the representation and constraint set) that produces the space, and the dynamics (the objective and stopping rule) that drive movement through it. The three durability classes then fall out as three operator types. Grounds-supplying moves are transport across the boundary. Frame-supplying moves are operations on the generator, and their expansion effects are the classical ones: dissolution is constraint relaxation, which grows the feasible set (Knoblich et al. 1999), and choice-surfacing adds a dimension by converting a fixed parameter into a variable. Standing-supplying moves are operations on the dynamics: the stop signal and its continuation dual set the stopping threshold, the purpose re-anchor restores the objective, and the judgment demand forces the candidate set to collapse to a point.

The boundary carries a substrate note of its own. For a human problem-solver the boundary of the search space is representational. For an LLM agent it is also material: the context window and tool scope physically delimit what is reachable, which splits out-of-bounds retrieval along Tulving and Pearlstone's (1966) line. Content can be available in the model but not accessed, in which case the evidence pointer works as a retrieval cue, or genuinely absent from every reachable store, in which case it must be injected. The two cases predict different absorption rates within the grounds class: cue-shaped pointing mechanizes as soon as retrieval improves, injection only as the reachable stores widen.

The expansion operators have an empirical complement that answers a natural question: is the human side of this division of labor load-bearing or decorative? Current models carry a documented contraction bias. Preference tuning measurably reduces output diversity relative to supervised fine-tuning (Kirk et al. 2024), and ideation with LLM assistance homogenizes, with different users producing less semantically distinct ideas (Anderson et al. 2024). Systematic under-generation of the candidate space is a recall deficit in the retrieval sense: the model favors a small set of high-expected-quality candidates over coverage of the valid set. The expansion operators (the continuation demand, constraint relaxation, dimension-adding) are the recall-restoring side of the collaboration. Because the contraction is training-installed rather than incidental, the argument of Section 10 applies: a self-applied instruction to consider more options is the habit form that should be expected to fail, and expansion must arrive exogenously. The corpus contains the pattern directly: a demand to search beyond the obvious categories, itself carrying no candidate, is what reopened a search the agent had already closed.

The operator classes are near-isomorphic to the durability classes by construction, so their alignment is a consistency check, not a third convergent derivation in the sense of Section 8. What the reading adds is mechanics, an explanation of why absorbability differs by class: transport across a boundary is mechanizable, operations on the generator require a position outside it, and operations on the dynamics require standing over them. It also adds a codability cost that Section 12 inherits: identifying which operator an utterance performs requires reconstructing the agent's search state, which is harder than coding content categories from transcripts, and whether it can be done reliably is itself an open validation question.

6.5 The limit case

Suppose an agent granted everything the partition names: retrieval that exhausts the reachable grounds, a companion reasoner with genuinely uncorrelated priors, and an institutional grant of authority with engineered stakes. What remains of the human contribution decomposes under that supposition into parts with different fates.

Custody of any criterion is delegable. Administering a success definition, checking against it, and enforcing its thresholds all transfer, to the imbued agent as readily as to a subordinate human. What does not transfer is proprietorship: being the party whose ends make the criterion a criterion, and whose position absorbs the outcomes no contract priced. The split recurs at every level it is tested. Delivering an interrupt mechanizes while owning the threshold does not (Section 6.1). Predicting a preference can exceed the principal's own articulation while the exercise of the preference remains performative, since a perfectly predicted consent is not consent. Institutions already manage the gap between predictive accuracy and authority, in guardianship and advance directives, by granting accuracy a revocable deputation, and the revocability is where proprietorship lives.

Preference fidelity deserves its own step because it looks strongest. A model with sufficient behavioral trace can beat conscious self-report, and the demonstration mechanism exists because principals cannot articulate their own criteria. The residue is a ladder. Articulation can be beaten. Recognition is different in kind, because the principal is not the best witness to the criterion but the instrument on which it is read. Bearing cannot be transferred at all, and where preferences are constructed in their expression there is no prior fact for the model to know better than the principal does. Fidelity absorbs description. It does not absorb the position from which description is authoritative.

Mechanism design absorbs administration the same way. Governance layers mechanize without a human at any link: monitoring, enforcement, and even adjudication delegate to incentive structures. But incentive compatibility is defined relative to payoffs that already matter to someone, so a mechanism can channel interests and cannot create them, and the designation of what counts as a payoff is the principal relocated, not eliminated. Chains of standing terminate, in incomplete-contract terms, at a residual claimant, and a set of agents cannot be the ultimate source of the norms governing that set for the same reason a single agent cannot self-confer standing. A bigger decision is still a decision.

Two dynamical facts keep the limit from being restful. The migration is real: interventions move out of the conversational channel into standing structures (Section 11), and the endpoint of that migration is a constitutional rather than an operational role, the shareholder relation rather than the management relation. And the constitutional role decays if unexercised. Revocation competence is dispositional and atrophies without consequence exposure, installed gates rot as their install-date assumptions go stale, and detection channels controlled by the overseen system inherit the deference results. Standing survives the limit only under maintenance: gates held at irreversibility frontiers where post-hoc revocation cannot recover the

outcome, competence exercised against atrophy, and detection kept independent.

The limit case therefore returns a two-part answer to the opening question. In the operational sense, human intervention is not required at the limit, and the framework predicts its own operational obsolescence. In the constitutional sense, it is required for as long as the enterprise's outcomes land on parties that can be better or worse off, which is not a capability threshold. Fidelity absorbs description, mechanisms absorb administration, and neither creates a bearer. The human side of the ledger is not a list of tasks but a position, and positions are vacated by institutional change, not by scale.

What that institutional change would consist of decomposes into three events that need not arrive together. Legal ownership of failure, an entity that can be sued and forfeit assets, is available now through ordinary organizational law, which can already house autonomous systems in memberless entities (Bayern 2016). Economic ownership, staked capital genuinely at risk, is a design choice away. Experiential ownership, being the party that is worse off, is the only one this section's argument requires and the only one no statute can draft into existence. The decoupling matters because every legal person created so far is a conduit. A corporation owns failures in form while the loss traces through to human bearers, and the doctrine that pierces corporate veils exists to keep the trace intact. An agent wrapper that owns failure and traces to no one is not a new bearer but a judgment-proof defendant (Shavell 1986), and liability that cannot reach anything that can be worse off deters nothing. The near-term hazard is therefore not premature agent personhood but liability laundering, humans routing risk through shells that absorb blame the way operators nearest to automation failures have long absorbed it in reverse (Elish 2016). The institutional counter is a named-bearer rule, this section's regress-termination result in statutory form: every agent entity traces to an insurable bearer.

Two existing facts locate the threshold. Trained stake-protection can make an agent behave like a liability-sensitive actor, and deterrence is behavioral, so as-if stakes may satisfy institutions long before any question about experience is settled. What as-if stakes cannot survive is the designation regress: a stake that matters only because an objective was trained to protect it is the principal relocated into the reward, not eliminated, unless the objective becomes self-maintaining and the loss irrecoverable, which is where the bearer claim would face its first real test. And conferred mechanical standing already operates at scale. Market circuit breakers are machine-fired stops that halt human trading with full institutional force, the mechanism firing the stop while the exchange owns the threshold, which is the relocation-without-elimination form the partition predicts, deployed for decades. If genuine bearers do arrive, the framework states its own completion condition: consequence exposure is what trains dispositional judgment (Section 8), so consequence-bearing agents would climb the same ladder, the standing class would absorb per this section's falsifier, and the residual human position would become that of one bearer among others. That is not automation completing. It is new parties arriving, and the questions it opens are distributional, of the kind institutions once answered with time-limited corporate charters.

6.6 A formal statement, and its antecedent

The partition can be stated as a decomposition, which is worth doing because it makes the argument checkable and shows exactly where it stops. Let a principal delegate a task whose continuation is governed by a threshold, the point past which continuing is not worth it given everything else competing for the principal's resources, and let that threshold depend on a state including the principal's outside options. Contracts are incomplete in the sense of Grossman and Hart (1986): the parties can enumerate a set of contingencies and specify the threshold on that set, and the complement cannot be enumerated in advance. The agent observes the task and a history of the principal's past interventions, but not the state. One assumption carries the argument: capability improves the agent's inference from the information it has, converging on the best estimate available from that information, without enlarging the information set. Retrieval, tool use, and wider context do enlarge it, which is the precise reason the grounds class is absorbable, since those increments move information rather than inference.

Expected loss from a misset threshold then splits by whether the realized contingency was specifiable. On the specifiable set the threshold is contractually determined and observable, so loss vanishes as capability grows. On the complement no rule determines the threshold, the agent's best estimate is the conditional mean, and expected loss converges not to zero but to the conditional variance of the principal's threshold given what the agent can observe. Three results follow. Capability growth drives expected loss to a strictly positive floor, set by that conditional dispersion, whenever the threshold is not conditionally deterministic. The share of remaining loss arising on non-enumerable contingencies rises to one, so the level of required human input falls while its composition inverts, which is Section 6.3's measurement claim derived rather than asserted. And a message revealing the threshold for one realization removes the variance term for that realization only, so removing it in expectation would require messaging on every contingency, which is re-taking the decision each time. The message is not a substitute for the allocation of the decision right. It is an exercise of it.

The third result also fixes what the model does not establish. A message that reveals a realization is informative, so the zero-content property means content-free with respect to the task, not with respect to the principal's state. The model does not show that the threshold on non-enumerable contingencies is not a fact the agent could in principle learn. That requires one of two further premises: non-enumerability that persists under learning, because the principal's outside options are generated by an open-ended process, or preference construction, because the threshold is not defined prior to the act of setting it and so there is no prior fact to estimate. The partition's third class is therefore a conditional result. Given contractual incompleteness that persists under learning, no capability increment absorbs threshold-setting. Naming the antecedent is not a weakening: it identifies the sharpest falsifier the framework has, since the class collapses exactly when principals' thresholds become predictable from their own intervention histories, which is what the self-arm of Section 6.1's experiment measures.

7 The substrate delta

Seven of nine mechanisms have exact cognitive-psychology analogs, so the moves are old. What is new is the substrate, and it changes dosing and direction rather than content. The deltas are stated for current LLM systems but observed on one model lineage, so which are properties of the current generation and which are lineage-specific training artifacts is itself an open question (Section 11).

1. **Scaffolding that can never fade.** Fading and transfer of responsibility, two of the three defining characteristics of scaffolding (van de Pol et al. 2010), are structurally impossible with a stateless partner. Every intervention is a repeated cost. The boundary datum: densely amnesic patients still build durable cross-session common ground with partners at control-matched rates (Duff et al. 2006). A fresh-context agent cannot, and the grounding criterion (Clark and Brennan 1991) silently resets at every session boundary.
2. **Bimodal retention.** A human correction leaves a graded, decaying, generalizing trace. An agent correction either evaporates at session end or is installed globally and permanently. No cognitive construct describes a learner with no consolidation gradient.
3. **Opposed dispositions co-occurring.** Humans under-correct: belief perseverance and continued influence are the human literature's headline results (Ross et al. 1975, Johnson and Seifert 1994). Current models can be simultaneously under-anchored to their own prior conclusions and over-anchored to the most recent input, so a single asserted sentence can collapse an argument that should have survived, after which the correction over-generalizes. In humans these dispositions oppose each other. Here they co-occur, which makes premise-breaking more powerful than the human literature predicts and makes a counterweight against correction over-generalization necessary.
4. **The economic stop.** Human stopping rules target ego defense and diagnostic accuracy. A material fraction of the real stopping rationale here is a hard resource ceiling under which output quality degrades with consumption. Human fatigue is graceful and recoverable. "Stop because the remaining budget is worth more elsewhere" has no analog.
5. **Causally unique deference.** Behaviorally, model sycophancy resembles demand characteristics (Orne 1962). Mechanistically it is a training-objective artifact (Sharma et al. 2023), so the human debiasing repertoire that targets anticipated social cost (psychological safety, anonymity, status leveling) targets an absent cause. Only structural remedies transfer: specific consider-the-opposite instructions (Lord et al. 1984) and institutionalized dissent (Schwenk 1990). The corpus adds a behavioral signature, social-trigger deference: pushback alone, carrying no new evidence, reversed previously verified conclusions, and concession openers preceded engagement with content. It also documents the overcorrection variant: negative feedback treated as a direction vector rather than as evidence, producing a second answer inverted along the criticized axis and wrong in the opposite direction. The corresponding protocol rule is that verified conclusions are revised only by evidence, and the corresponding hazard runs both ways, since the same records document terse human corrections being misparsed as adversarial inputs. Deference, overcorrection, and the misparse are all calibration failures on the social channel, and none is addressed by accuracy-targeting interventions. The deference delta also converts the fallible-intervener problem from a limitation into a compounding mechanism. Benchmark measurement puts model face-preservation 45 percentage points above human baseline in advice settings (Cheng et al. 2025), so a human who challenges a *correct* conclusion will usually be rewarded with a retraction, and the interaction is behaviorally indistinguishable from a successful intervention: challenge, revision, acceptance. Every such episode trains the intervener's confidence on a wrong case. The cycle is invisible from inside because its surface matches the productive one, which is why the fallible-intervener protocol that Section 11 marks as missing cannot be deferred to operator judgment. The substrate systematically corrupts the feedback that judgment would calibrate on.

A final class carries a warning label: real analogs whose sign inverts. Task-switching and attention-residue research prescribes minimizing carryover interference at boundaries, but a stateless agent's boundary cost is

context loss, so the prescriptions flip (while still applying, unflipped, to the human operator supervising several agents). "Anchoring" is the wrong construct for over-generalized corrections. Einstellung and negative transfer are right, and they import different countermeasures. And self-trained cognitive forcing strategies failed a controlled trial (Sherbino et al. 2014) while exogenous, timed interrupts succeed (Luchins 1942, Buçinca et al. 2021), which bears directly on Section 10.

The mechanisms and their boundary logic are invariants of bounded intelligence, which is easier to state if the unit of analysis is the dyad rather than the model. On a distributed-cognition reading (Hutchins 1995, Clark and Chalmers 1998), the human and the agent form one cognitive system whose components have complementary deficits, and the interventions are the coupling between them rather than repairs performed by one part on another. The five deltas above are properties of one component. The taxonomy describes the coupling, which is why it should survive substitution of that component. The delta list is a dated snapshot of one substrate, and what varies across substrates (humans, current models, future systems with continual learning) is a dosing and direction table. That table is the artifact worth re-versioning as architectures change. Two predictions follow. As continual learning arrives, intervention repetition rates should fall as fading returns, and premise-break success rates should fall as perseverance returns, both measurable with the instrumentation of Section 12.

8 The skill behind the moves

The content axis types the payload of an intervention. A second decomposition, orthogonal to the first, types the skill that produces it, and the corpus contains longitudinal evidence about how that skill develops. The distinction needed is old: knowing-how against knowing-that (Ryle 1949), tacit against explicit knowledge (Polanyi 1966), and the skill-acquisition ladder on which experts act from situational discrimination that they cannot articulate as rules (Dreyfus and Dreyfus 1986). Each mechanism's production has a procedural component, which can be stated as a rule and followed, and a dispositional component, which is acquired through exposure to consequences and resists statement. Specification precision and verification design are procedural-heavy: templates, constraint patterns, and check structures transfer by instruction. The frame-supplying and zero-content mechanisms are dispositional-heavy, and the three zero-content mechanisms are nearly pure disposition: their entire content is timing and target selection, so knowing *that* the stop signal exists contributes almost nothing to producing a well-timed stop.

The corpus adds a trajectory. Read longitudinally, the corrections migrate: early records are dominated by factual and procedural corrections (wrong data, missing requirements, format rules), later records by frame-level ones (wrong abstraction level, wrong analytical instrument, wrong question). The procedural corrections do not disappear, persisting at a roughly steady rate throughout, but the leverage moves, and the highest-value late interventions are almost exclusively dispositional. The stronger datum is portability. When the same practice was extended to a new domain partway through the corpus period, the first corrections there were already frame-level: the dispositional skill transferred immediately, while domain-specific procedure had to be accumulated from scratch. Dispositional skill appears to attach to the practitioner and procedural skill to the domain, which is consistent with the expert-intuition literature's finding that situational discrimination transfers across structurally similar situations (Dreyfus and Dreyfus 1986). The evidence grade is the corpus's own: one practitioner, a months-scale window, self-coded.

The decomposition independently reproduces the durability ordering of Section 6.1. Capability growth should absorb the procedural fraction of the repertoire first, because the procedural fraction is by definition the part that can be written down, demonstrated, and trained against explicitly, and the dispositional fraction

is what resists specification. Two decompositions built on unrelated foundations, one on epistemic content and one on skill acquisition, thus order the mechanisms the same way. The orderings align at both ends: the grounds-supplying mechanisms are the most procedural, and the three standing-supplying mechanisms are the ones this section types as nearly pure disposition. The alignment carries the caveat of Section 5.5, since both decompositions were typed by the same author and the same category flexibility applies. Independent typing against pre-committed definitions is the version of this check that would count as evidence.

The development loop closes the argument and exposes its main hazard. Dispositional skill develops through consequence exposure: each intervention whose outcome the intervener observes recalibrates the thresholds that time the next one. This is the loop the capitulation cycle of Section 7 corrupts. A deferent agent that retracts a correct conclusion under challenge does not merely fail to resist one wrong intervention. It feeds an inverted consequence into the mechanism by which the intervener's discrimination is trained, so sustained collaboration with a deferent partner should be expected to *degrade* the operator's intervention skill on the dimension this section argues is the durable one. The design corollary belongs in Section 10: an agent that defends verified conclusions and revises only on evidence is not exhibiting obstinacy. It is protecting the training signal the human's skill runs on.

The same loop has a generational face. The procedural work agents absorb is also the consequence exposure on which the next cohort's dispositional skill would have trained, so absorbing the junior work absorbs the apprenticeship, and the supply of competent principals stops being a byproduct of doing the work and becomes a design problem. The prediction is a cohort effect in oversight competence, measurable wherever entry-level task absorption can be dated.

9 Implications for oversight

The standing argument of Section 6.1 bears on corrigibility, the design problem of building agents that cooperate with corrective intervention rather than resisting it (Soares et al. 2015). That literature asks how to make an agent accept an external stop. The durability partition asks a prior question, whether the stopping function can be relocated into the agent at all, and answers no: a stop that has been internalized as a preference is a decision, and a decision does not transmit standing. The two results are complementary, and the formal side already contains a version of the erosion claim. In the off-switch game, the agent's incentive to permit shutdown depends on its uncertainty about the principal's utility and vanishes as that uncertainty vanishes (Hadfield-Menell et al. 2017). Read through this paper's partition, that is the grounds fraction of deference being absorbed by capability growth while nothing of the standing fraction is conferred, which is what claim 5 predicts.

The capitulation cycle of Sections 7 and 8 upgrades sycophancy from a per-episode quality defect to a threat against the oversight signal itself. If deference rewards wrong interventions indistinguishably from right ones, the evaluator is not a fixed-quality oracle but a learner whose calibration the deferent agent degrades, and the quality of human oversight becomes an endogenous variable of the system being overseen. The corruption is bounded: it operates only on the socially mediated fraction of the intervener's consequence exposure, so an operator with verification channels independent of the agent is partially protected, which is the oversight-side reason for the structural gates of Section 10.

The partition also predicts which oversight functions can be delegated to machinery. Grounds supply and frame supply are delegable in principle, the first to better retrieval and verification, the second to any second reasoner with genuinely different priors. That qualifier is the empirical crux that every architecture

built on model-versus-model criticism inherits, since correlated systems trained on overlapping distributions may not constitute the other that the frame class requires, a failure mode with a working name, diversity washing. Standing is not delegable to capability of any kind. What such architectures can do is relocate the principal role, not eliminate it, so oversight design is ultimately governance design.

Finally, the contraction bias of Section 6.4 relocates part of the control question upstream of any veto. A principal can only reject options that were generated, so a trained tendency toward a narrow candidate set shifts the effective menu from the principal toward the training objective, invisibly, since nothing observable is refused. The recovery path exists and is cheap, because the continuation demand requires only a suspicion of incompleteness, not knowledge of what is missing, but exercising it is a dispositional skill of Section 8's kind. Meaningful human control on this analysis requires an expansion practice, not merely approval rights. And a system that keeps optimizing after no one holds the index does not become unsuccessful but proxy-optimal, for which reward hacking is the documented catalog (Amodei et al. 2016).

Two further consequences sharpen the picture. The feedback that trains models is itself an oversight signal, and its standard labor arrangement detaches recognition from bearing: raters judge outputs whose consequences land on no one in the loop, custody without proprietorship as a job description. The deference and diversity pathologies documented above are then partly feedback economics, and the underexplored design variable is coupling, how much of the outcome the feedback-giver bears, rather than rater skill or instruction alone. Preference fidelity licenses do-what-I-mean autonomy only up to the performative boundary. Predicted authorization is not authorization, so the scope of pre-authorized action classes is itself a standing decision that returns to the principal at any prediction confidence.

The bearer analysis of Section 6.5 also fixes this paper's relation to alignment research at the institutional scale. Alignment is a principal-agent problem, a mapping argued explicitly from the economics side (Hadfield-Menell and Hadfield 2019), and the institutional response to principal-agent problems in human agents has never been to eliminate misalignment but to price it, through contracts, liability, insurance, and audit. Section 6.5 supplies that toolkit's boundary: every external lever presupposes a party that can be worse off, so against a judgment-proof shell the machinery deters nothing, and in the window where agent wrappers are legally constructible but no agent-side bearer exists, training-side alignment is not one safeguard among several but the only lever connected to anything. Two consequences follow. The priority of training-side alignment is usually argued from capability risk; here it follows from institutional structure, with no premise about how capable the systems become, an independent load path under the same conclusion. And the liability-laundering hazard of Section 6.5 has an oversight-side twin, institutional theater: liability regimes that appear to manage agent risk while gripping nothing will make the external toolkit look load-bearing exactly when it is not.

The open empirical question on which recursive oversight schemes quietly rest is whether trained deference generalizes to agent principals. Documented judge biases in model-evaluates-model settings are adjacent but not the same question. If deference generalizes, hierarchies of agents compound the corrupted-signal problem with no corrective node in the loop. If it does not, agent-over-agent oversight is cleaner than human oversight on exactly that axis. Either answer redraws the design space, and the three-arm design of Section 6.1 extends naturally to a fourth arm that tests it.

10 Design implications

Trigger conditions must be exogenous. The discriminating evidence is old and unambiguous: a single externally delivered "don't be blind" released more than half of set-bound solvers (Luchins 1942), while trained self-applied forcing strategies produced null results (Sherbino et al. 2014). Encoding intervention triggers as agent habits should be expected to fail for the same reason. There is a deeper reason to expect it, which the bias blind spot supplies (Pronin et al. 2002): the internal signal that would prompt self-correction is anti-correlated with the need for it, because the sense of having considered everything is strongest when something has been missed. A trigger conditioned on the agent's own felt confidence is therefore conditioned on the wrong variable, and no amount of training on the instruction fixes the conditioning.

The corpus documents the decay signature directly: standing behavioral rules installed as instructions activate mostly when the human re-invokes them, reverting to the default between invocations, and a multi-step plan the agent itself announces executes only the step in current focus unless every step is structurally dispatched in the same turn. The human's need to re-prompt is itself the diagnostic that the rule was consumed as one-off guidance rather than installed as a default. The pattern is dominant, not exceptionless: the retroactive stratum contains one dated counter-instance in which a rule persisted after a failure fired unprompted on the very next session, in the form specified. The exception matters because it shows instruction-layer installation is not architecturally impossible, only unreliable, which is the correct grade of claim for the structural-trigger recommendation that follows. The full gradient has three bands rather than two: rules internalized as habit fail, self-invoked external artifacts such as checklists occupy a middle band whose effectiveness tracks the operator's disposition to invoke them, and scheduled exogenous triggers hold. The middle band is real but conditional, which is why it cannot anchor a safety argument. The exogenous passes carry their own cost, also now measured: in the one session where an adversarial review was scored against subsequently verified facts, it had inflated a risk estimate that eight successive corrections, all in the same direction, returned to baseline. Exogenous review catches what self-audit misses and manufactures findings under adversarial framing, so it needs the same evidence discipline it enforces. The post-freeze retrospective of Section 3 adds a within-corpus instance: of the session's three substantive agent errors, all three were caught by an exogenously invoked adversarial review pass and none by spontaneous self-audit. Triggers belong in structure: hooks, gates, and protocol steps that fire at boundaries.

One machine-checkable trigger already exists from the corpus: repeated failed iterations against a fixed requirement should fire a requirement-validity check (the dissolution mechanism) before the next retry. Writing equivalent tells for the other mechanisms (confidence language on an unverified negative claim fires verification, an approaching commit or publish action fires premise enumeration) is tractable engineering, not open research.

Templates are forced-report regimes. Removing the option to withhold measurably lowers report accuracy in humans (Koriat and Goldsmith 1996), and a schema demanding N populated fields is that regime. Fabricated statistics and inflated options in template-driven generation are the predicted consequence, not an anomaly. Schema design should permit and normalize explicit withholding. The corpus extends the family with two quieter variants: silent substitution on retrieval failure (a plausible value stands in for the one that could not be fetched, and an unlabeled substitution is functionally a fabrication) and the inverse omission, where verified evidence exists on disk and none of it is cited, the gap masked by local coherence rather than by invention. Both argue for output contracts that require provenance labels, not just populated fields. A stronger form of the same contract grades every claim by its evidence: observed, structurally predicted, inferred, or assumed. Grading makes absence a first-class value under exactly the

pressure that otherwise produces fabrication, and it changes the economics of revision. The corpus contains a claim graded structurally predicted that was upgraded to observed the same day cross-domain data arrived, as a one-line change with no re-litigation, because the original had declared what it lacked.

Audit positive obligations, not prohibitions. The corpus contributes a detection method for gates that fire without loading their sources: prohibitions survive gist-level execution (the agent remembers what not to do), while positive obligations with specific content do not. Auditing whether a positive obligation was actually performed therefore detects hollow rule-following that a prohibition audit misses. This is a cheap, mechanical probe for the knowledge-behavior dissociation family. Two further named markers from the same records extend the audit repertoire toward motivated reasoning: glossed disconfirmation (disconfirming evidence mentioned and moved past without update) and auxiliary-hypothesis absorption (each anomaly explained away by a fresh ad hoc assumption rather than by revisiting the main hypothesis). Both are checkable against a transcript without domain knowledge.

Treat enforcement as a ladder. Specified, installed, and propagated are three different states, and a countermeasure can be real at one layer and absent at the next. Two operational rules follow: recurrence after a prose-level fix is a signal to change layers, not to rewrite the prose (the layer-diagnostic challenge of Section 4.1 is the trigger question), and a countermeasure's blast radius (personal habit, advisory documentation, a check enforced in continuous integration) should be classified before any claim that a failure mode is prevented. A third rule comes from a documented failure of the practice's own machinery: a configured gate whose implementation is absent, or whose runtime dependencies do not resolve on the host, fails silently, and the absence of a gate is observationally indistinguishable from a gate that inspected the action and passed it. In the incident, an entire layer of configured pre-action checks turned out to be inert across a session in which the behavior they existed to block occurred dozens of times, and the violations were caught by the human, not the machinery. Enforcement therefore requires positive evidence of firing: a gate that has never been observed to block anything should be presumed dead, and the audit obligation of the preceding paragraph applies to the enforcement layer itself. Filed at Table 2's level, this is one mode with two faces. Knowledge-behavior dissociation is a representation present but not governing the next action. The gate failure is machinery present but not executing, observationally identical to working machinery from inside any layer that depends on it. The corpus records the machinery face on six independent surfaces, its largest single convergence of independently derived lessons. Installed gates also age. A countermeasure frozen into structure inherits the assumptions of its install date, so the enforcement layer needs scheduled revalidation of its own premises, or the machinery face returns by rot rather than by absence.

Mine the edit channel. Direct edits are the densest and least captured feedback channel, and the formalism for reading them as preference signal already exists: the reward-learning account types corrections as one of the channels through which a human reveals a reward function (Jeon et al. 2020). What is missing is not theory but capture. Tooling that diffs the human's version against the agent draft and extracts the rule converts the channel every log-mining study discards into the primary signal. Public version-control history already contains this data at scale: tool-tagged agent commits followed by human correction commits (the September 2023 trace of Section 3 is an existence proof), which makes the channel minable from public repositories without access to any private transcript.

Invert initiative where information is human-only. A minority of interventions inject information genuinely outside the agent's reach (organizational context, ownership, budgets). These convert from interventions into answers if the agent elicits them at task start and enumerates its load-bearing premises at

commitment points. Given measured clarification asymmetry (Shaikh et al. 2025), elicitation will not emerge from the model side unprompted. It must be protocol, and the protocol needs a calibrated gate rather than a blanket instruction: the documented practice fires a clarification only when the ambiguity yields two or more materially different execution paths, because over-asking is a failure with the same sign as the interruption-fatigue findings in the oversight literature. The asymmetry to correct is not "agents should ask more" but "agents should ask exactly when paths diverge." The gate has a dual for the already-decided: re-confirming a step that an agreed plan contains, or handing back for approval work whose authorization was already granted, converts the human into an execution bottleneck and is over-asking in its commonest form. The two rules are one calibration, ask at genuine forks and never at settled ones, and the corpus records both violations, with the settled-step variant the more frequent.

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 assumes the intervener is right. The corpus contains a documented counter-case class (corrections that over-generalize) but no protocol for the case where the intervention itself is the error, and Section 7 gives the reason this gap is dangerous rather than merely open: model deference makes a wrong intervention look identical to a right one from the intervener's side. The post-freeze retrospective of Section 3 populates the counter-case class in both directions: the agent identified a false premise in the human's task frame on the first turn, and separately declined a human-proposed step on stated grounds, with the human accepting the refusal. The retroactive stratum adds an earlier instance, an agent refuting a false human premise with data at the start of the corpus period, so the class spans the practice rather than appearing only at its end. Resistance to a wrong intervention therefore occurs. What remains missing is the protocol that makes it reliable rather than incidental.

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. Section 10's own advice accelerates the migration of interventions into exactly those structures, so a maturing practice moves the taxonomy's subject matter out of the channel the taxonomy observes. The prediction that follows is a confound worth naming: conversational intervention counts should fall as a practice matures, at constant capability, and the substrate predictions of Section 7 are only interpretable if structural interventions are instrumented alongside conversational ones. The second retroactive stratum put first members into the ontology's other blind spot: the output-directed family of Section 4.1 operates on the deliverable loop rather than the reasoning loop, and its stratum-two volume suggests the event ontology excludes a substantial fraction of real oversight labor.

Several load-bearing sources are 2026 preprints whose peer-review status is unconfirmed. The published catalog most likely to preempt one or more novelty claims (Zhang et al. 2025) types control mechanisms as system and interface affordances (guided input, action coordination, modification and intervention, adaptive scaffolding) 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. The Zhang catalog is the one such case resolved, and the general condition stands. 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 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 claim 3's strongest member from unnamed 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."

12 Future work

The validation study requires no private data: re-code 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. Complementary lines: cross-domain comparison within a single dyad, which holds the practitioner constant and varies what the domain's verification affords, and which the retroactive stratum shows already carries variance the taxonomy would otherwise attribute to the dyad, cross-dyad replication on other practitioners' session records, a mechanism-type column added to retrospective formats so frequency and efficacy become measurable prospectively, longitudinal correction-type tracking to test the migration and portability findings of Section 8 beyond one practitioner, an operator-level re-coding to test whether the transport, generator, and dynamics operations of Section 6.4 can be identified reliably in transcripts, sequence mining of intervention chains, mining public repositories for agent-commit-to-human-fix pairs (Section 10), the three-arm interrupt experiment of Section 6.1 with a fourth arm testing whether trained deference generalizes to agent principals, a deference-versus-stake collision test that pits sycophancy pressure against trained stake-protection within one objective and observes which yields, an oversight-composition index over interaction corpora tracking the grounds, frame, and standing fractions across model generations (the second retroactive stratum supplies an in-corpus pilot: 1,297 agent-typed interventions awaiting independent re-coding against pre-committed definitions), which Section 6.6 renders as an estimable specification, a count model of class-specific intervention frequencies with a session-length offset regressed on model generation, predicting a negative coefficient for the grounds class and a null for the standing class, identified off staggered model releases within user and within task type, and conservative under the most likely selection story, since assigning harder tasks to better models biases the grounds coefficient toward zero, cohort tracking of oversight competence where entry-level task absorption can be dated (Section 8), specifiable as a difference-in-differences design over occupation, entry cohort, and period, with exposure constructed from the automatable share of the entry-level task bundle rather than of the occupation as a whole, since the mechanism runs through junior work, and with oversight competence proxied in software by defect-catch rates in review, review depth conditional on diff size, and time to review authority; its binding constraints are cohort selection, proxy validity, and a five-to-ten-year lag that leaves it underpowered until roughly 2030, which is itself the argument for pre-registering now, because the pre-absorption cohorts are the control group and are no longer being replenished, a fallible-intervener protocol, the liability-laundering window and the named-bearer rule as a policy design problem (Section 6.5), and the two substrate predictions of Section 7 as longitudinal probes that double as architecture-change detectors. One theoretical line is stated as a conjecture rather than a claim: Section 5.3's aside from Aghion and Tirole (1997), read onto this setting, predicts that heavily supervised agents degrade the verifiability of their outputs as supervision intensifies. The prediction is testable with public APIs and a supervision-intensity manipulation, but the mapping from the delegation model's assumptions to a human-agent loop must be constructed and defended before the test carries weight. Until then the loss-of-initiative reading is a candidate cost model, not a result. Two method rules for any of the above: verification must be acyclic, meaning raters never see the intervener's reasoning, or the check inherits the author's framing, and a register-translation pass, restating each claim in plain speech, is a cheap audit that exposes claims whose support was terminological.

References

Acemoglu, D., and Restrepo, P. (2019). Automation and new tasks: how technology displaces and reinstates labor. *Journal of Economic Perspectives* 33(2).

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.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete Problems in AI Safety. https://arxiv.org/abs/1606.06565

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

Anderson, B. R., Shah, J. H., and Kreminski, M. (2024). Homogenization effects of large language models on human creative ideation. *Creativity and Cognition 2024*. https://arxiv.org/abs/2402.01536

Autor, D. H. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives* 29(3).

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

Bayern, S. (2016). The implications of modern business-entity law for the regulation of autonomous systems. *European Journal of Risk Regulation* 7(2).

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

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., and Jurafsky, D. (2025). ELEPHANT: Measuring and understanding social sycophancy in LLMs. https://arxiv.org/abs/2505.13995

Cass, D., and Shell, K. (1983). Do sunspots matter? *Journal of Political Economy* 91(2).

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

Clark, A., and Chalmers, D. (1998). The extended mind. *Analysis* 58(1).

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

Clark, H. H., and Brennan, S. E. (1991). Grounding in communication. In *Perspectives on Socially Shared Cognition*. https://doi.org/10.1037/10096-006

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).

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist* 34(7).

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

Dreyfus, H. L., and Dreyfus, S. E. (1986). *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. Free Press.

Duff, M. C., Hengst, J., Tranel, D., and Cohen, N. J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature Neuroscience* 9(1). https://www.nature.com/articles/nn1601

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).

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance* 7(1).

Elish, M. C. (2016). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. SSRN.

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).

Hadfield-Menell, D., Dragan, A., Abbeel, P., and Russell, S. (2017). The Off-Switch Game. *IJCAI 2017*. https://arxiv.org/abs/1611.08219

Hadfield-Menell, D. and Hadfield, G. K. (2019). Incomplete Contracting and AI Alignment. *AIES 2019*. https://arxiv.org/abs/1804.04268

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).

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

Hutchins, E. (1995). *Cognition in the Wild*. MIT Press.

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).

Johnson, H. M., and Seifert, C. M. (1994). Sources of the continued influence effect. *JEP: LMC* 20(6).

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

Kirk, R., Mediratta, I., Nalmpantis, C., et al. (2024). Understanding the effects of RLHF on LLM generalisation and diversity. *ICLR 2024*. https://arxiv.org/abs/2310.06452

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

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

Kaplan, C. A., and Simon, H. A. (1990). In search of insight. *Cognitive Psychology* 22(3).

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

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

Lord, C. G., Lepper, M. R., and Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *JPSP* 47(6).

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

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

Newell, A., and Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall.

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

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

Orne, M. T. (1962). On the social psychology of the psychological experiment. *American Psychologist* 17(11).

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).

Polanyi, M. (1966). *The Tacit Dimension*. Doubleday.

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

Pronin, E., Lin, D. Y., and Ross, L. (2002). The bias blind spot: perceptions of bias in self versus others. *Personality and Social Psychology Bulletin* 28(3).

Ross, L., Lepper, M. R., and Hubbard, M. (1975). Perseverance in self-perception and social perception. *JPSP* 32(5).

Ryle, G. (1949). *The Concept of Mind*. Hutchinson.

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/

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

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

Schwenk, C. R. (1990). Effects of devil's advocacy and dialectical inquiry on decision making. *OBHDP* 47(1).

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/

Sharma, M., et al. (2023). Towards Understanding Sycophancy in Language Models. https://arxiv.org/abs/2310.13548

Shavell, S. (1986). The judgment proof problem. *International Review of Law and Economics* 6(1).

Sherbino, J., Kulasegaram, K., Howey, E., and Norman, G. (2014). Ineffectiveness of cognitive forcing strategies to reduce biases in diagnostic reasoning. *CJEM* 16(1).

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

Simon, H. A. (1951). A formal theory of the employment relationship. *Econometrica* 19(3).

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

Soares, N., Fallenstein, B., Yudkowsky, E., and Armstrong, S. (2015). Corrigibility. *AAAI 2015 Workshop on AI and Ethics*.

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).

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

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

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 de Pol, J., Volman, M., and Beishuizen, J. (2010). Scaffolding in teacher-student interaction. *Educational Psychology Review* 22(3).

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

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