Human Intervention in Agentic Work: A Measurement Scheme, Pilot Evidence, and a Research Program

Jongsun Suh

Independent Researcher

Preprint · 1 August 2026 · Research proposal v0.6

Research proposal

A measurement scheme, pilot evidence, and a research program

Jongsun Suh · 1 August 2026 · Research proposal v0.6

This is the proposal-length treatment. The technical record is two papers, split on 31 July 2026 along the dependency between the taxonomy and its entailments. *An Epistemic-Content Taxonomy of Human Intervention in Agentic Collaboration* carries the complete mechanism inventory, the failure-mode taxonomy, the discovery corpus, the coding scheme with its reliability measurements, the survey, and every novelty claim (concept DOI 10.5281/zenodo.21730478), and is cited here as Paper A. *Grounds, Frames, and Standing* carries the durability partition, the formal appendix, and the oversight and design implications (concept DOI 10.5281/zenodo.21563154), and is cited here as Paper B. Where this document compresses, those two are the record.

ABSTRACT

When, if ever, is human intervention required for an AI agent to operate successfully? The question is empirical and the field cannot currently answer it, because the available record registers every correction as one undifferentiated event. That matters at a scale already measured: in the largest published analysis of agentic coding, 91.49% of failure resolutions required explicit user correction against 2.99% agent self-correction. Frequency on its own says nothing about which of those corrections capability growth can remove. The proposal is that interventions be indexed by epistemic content rather than by timing, authority, or defect class, along two dimensions: what a move supplies, and what it operates on. What a move supplies sorts into three classes with different fates. Grounds, evidence bearing on entitlement to a conclusion, are absorbable by capability growth. Frames, redefinitions of the hypothesis space, yield to a second reasoner who need not be human. Standing cannot be self-conferred, so no capability increment absorbs it. Under incomplete contracting, expected loss from a misset delegation threshold converges to a strictly positive floor. The policy-facing corollary is a compositional prediction: the grounds-shaped fraction of human input declines as capability grows while the standing-shaped fraction does not, so task-level exposure indices can report stability across exactly the period in which the character of human work inverts. The decline carries a floor of its own, since absorbing a grounds move means absorbing its verification, and a verification that terminates in an act of standing rather than in a closed check absorbs down to that terminal act and no further. Pilot evidence from one practitioner's seven-month record supports the scheme's usability and does not establish its findings. Blind rater pairs drawn from different model tiers agree at kappa 0.77 on whether an input is an intervention at all and 0.70 on its family; on sessions held out from the scheme's development the gate replicates at 0.82 while family agreement falls to 0.58. One compositional result survives every protocol and replicates out of

sample: deliverable-directed correction is the largest family of human interventions, which is a blind spot of the event ontology the field currently uses. What the pilot cannot do, no existing instrument can do either. The field's flagship measurement partitions conversations by division of labour, which is orthogonal to content, and its sensitivity to a class runs inversely to that class's verbosity, making it structurally least able to detect the one class this argument says does not absorb. The program proposed here builds the message-granular instrument that gap requires, and runs the experiments that would falsify the partition, beginning with the one the partition names against itself.

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

1 The question, and why the record cannot answer it

Ask when human intervention is required for an agent to succeed, and the honest answer today is that nobody knows, because the question has not been asked in a form the data can answer. Interaction records count corrections. They do not distinguish a correction that supplies a missing fact from one that replaces the problem statement from one that carries no information at all and simply ends the inquiry. Those three have different futures. The first is the kind of thing better retrieval removes. The second needs someone standing outside the reasoning. The third is not information at all, and improving a reasoner does not touch it.

This proposal advances seven claims, of which the first three are conceptual, the next two methodological, and the last two are the ones a program would test.

Claim 1. Human interventions in agentic work sort by epistemic content into grounds, frames, and standing, and the sort is exhaustive over the live-reasoning tier, which is disciplined by the corpus rather than derived and carries the corpus's evidence grade.

Claim 2. The three classes have different absorption routes. Grounds are absorbed by capability growth, down to the depth at which the domain's verification bottoms out in a closed check rather than in an act of standing. Frames are absorbed by architecture, specifically by a second reasoner with different priors. Standing is absorbed by neither, because what it transfers is a threshold the executor does not own.

Claim 3. The supply of frames is contingent on a population rather than a capability, which makes the second claim's middle term conditional in a way the literature has not stated.

Claim 4. Content is a second axis, not a refinement of the existing ones. Timing, authority, granularity, and defect class are all orthogonal to what a move supplies, and the existing taxonomies index on those.

Claim 5. The axis is measurable. Blind raters applying pinned definitions agree well enough on the gate for the scheme to be operable, and less well on family assignment, and the honest report of that gap is part of the contribution.

Claim 6. The composition inverts as capability grows: the grounds fraction falls, to its verification floor rather than to zero, and the standing fraction does not, so aggregate exposure measures understate the change in the character of human work.

Claim 7. The partition names its own 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.

Claims 1 through 4 are argued here and in the two papers, the axis and its orthogonality in Paper A and the absorption routes in Paper B. Claim 5 has pilot evidence, reported with its failures. Claims 6 and 7 are what the proposed program exists to settle, and Section 8 states the designs, their predictions, and the conditions under which each would refute the framework.

2 Where this sits

Four literatures are adjacent and none occupies this position.

Human-AI interaction and oversight research indexes intervention by timing and authority: when a human may intervene, and with what force. Work on AI-assisted programming and agentic coding supplies the frequency data this proposal builds on, including the correction asymmetry quoted above, and types failures by defect class rather than by what the human contributed. Older disciplines already name many of the individual moves. Of the thirteen mechanisms in the full inventory, twelve carry established names across eleven literatures, which is a finding about vocabulary rather than about novelty: the moves are known, and no field indexes them by what they supply.

The fourth adjacency is the informative one. Cognitive and educational research builds extensive machinery for making humans think harder, and none of it studies the human as the forcing function on a machine's live reasoning. The direction is inverted, and the inversion is why the vocabulary does not transfer cleanly even where the moves match.

One convergence from a different tradition is worth recording rather than leaning on. An argument from mechanism design reaches a similar conclusion, that systems should be steered by sparse high-quality human input rather than granted maximal autonomy (Buterin 2025), on premises this proposal does not share. Convergence from disjoint premises is evidence about a conclusion's robustness and no evidence at all about its mechanism.

3 The scheme

3.1 Nine mechanisms, and what three of them lack

From seven months of one practitioner's records, nine mechanisms recur, with a lower-evidence second tier of four more. Three of the nine carry no task-domain content. A stop signal, a judgment demand, and a purpose re-anchor move commitment, salience, or attention, and transfer no proposition about the work. They are not uninformative. A stop is timed and it is signed, and both reveal a threshold's value. But a value is a fact, conveyable in advance and learnable from a record, so the residue is not information of another kind. What does not reduce to a fact is the authority to set and revise the threshold. The distinction is therefore what kind of act the message is rather than where its information sits. A stop is the exercise of a right, and its timing and sign are byproducts of that exercise rather than its function. The two come apart cleanly: verdicts are predictable at high accuracy, and predicting one is not rendering one. An agent issuing a stop to itself has either estimated the value, which is grounds about the principal and falls under the predictability component, or assumed the authority, which is a different act.

The remaining six divide by what they do supply. A premise-breaking fact, a cross-context evidence pointer, and a demonstration by direct edit supply grounds. Causal-model injection, dissolution of a requirement, and implicit-choice surfacing supply frames.

3.2 The second axis: what a move operates on

Content alone under-describes the record. The same supply can be directed at the live reasoning, at the produced deliverable, at the rule store governing later sessions, at the channel itself, or at the agent's self-model. Crossing supply with target produces the scheme actually used for coding.

The target axis was not designed. It was forced by the data: in a verbatim sweep of a second workstation's record, 178 of roughly 1,300 typed interventions, 14 percent, failed to map onto the mechanism inventory, and the dominant unmapped classes were demands that reject assertion in place of verifiable evidence, and callouts of defects in the agent's just-delivered output. The second class does not operate on live reasoning at all. It operates on an artifact.

The strongest objection to the target axis is that correcting an artifact is correcting the reasoning behind it, so the distinction collapses. It does not, for a reason the record shows directly. The two come apart in both directions. An artifact can be corrected with no change to the reasoning that produced it, as when a deliverable is edited by hand and the agent never learns why. And reasoning can be corrected with no artifact in existence yet. What the objection correctly identifies is that *targeting* the deliverable and *affecting* the reasoning are different questions, and the scheme codes the first, because the first is observable in a transcript and the second is an inference about effect.

4 Pilot evidence, and what it does not show

4.1 The corpus is a discovery instrument

The corpus is seven months of my own practice as a senior engineer working with LLM coding agents on a large production codebase. Twenty-six structured session retrospectives exist; fifteen entered the analysis. They were written against roughly 260 recorded working sessions, so coverage is about one session in fifteen.

Every property that matters for interpretation cuts against treating this as evidence for the findings. It is one practitioner, one agent lineage, self-coded by the collaboration under study. Retrospectives are written when a session produced something worth recording, which is disproportionately when intervention was needed and worked, so the corpus measures the composition of intervention where intervention happened and is silent on frequency and on sessions needing none. No frequency claim rests on it.

Two failures are worth reporting because they bound what any log-mining study can recover. Coding from conversation logs alone produced the finding that interventions rarely edit the artifact, inferred from the absence of edits in the logs. The finding is false. Agent-drafted artifacts are routinely revised by hand in ways that change their content, and that revision leaves no conversational trace. The transcript misled the analyst who had personally performed the missing edits, which sets an upper bound on what a third party can recover from the same record. Where edit deltas were captured they were the densest data in the corpus: one session's hand-edits yielded over a dozen distinct rules of style and register.

4.2 Reliability, reported with the drop

The full 2,205-input record was re-coded by blind rater pairs drawn from different model tiers, chosen because same-family raters inflate agreement through shared priors. With labelling conventions unpinned, family-level agreement is kappa 0.59. With vocabulary and boundary conventions pinned in the brief, the same corpus reaches 0.70 and the task-versus-intervention gate closes to 0.77. A same-family pilot's 0.86 stands as an optimistic upper bound.

The scheme was then checked against sessions it had never seen. A census of 105 session transcripts found 77 not represented in the corpus. An exclusion rule for research-about-the-research, applied on session topic before any coding, removed 986 of 1,083 candidate turns, leaving 93 turns across 10 sessions, 8.6 percent of typed input. That yield is itself a finding worth stating plainly: the practitioner's recent record has been consumed by the research about the record, and no further stratum is available from this source.

Axis	Out-of-sample (n=93)	In-sample (n=2,205)
Gate	87.1% agreement, kappa 0.818	kappa 0.771
Class	80.6% agreement, kappa 0.580	kappa 0.615
Family	80.6% agreement, kappa 0.575	kappa 0.699

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

The gate replicates and slightly improves; family agreement falls from 0.70 to 0.58. The task-versus-intervention boundary transfers to unseen sessions. Family assignment transfers less well. Both figures carry wide uncertainty at this n, so the direction is the finding and the magnitude is not. A scheme whose gate is reliable and whose families are moderate is usable for the compositional question and not yet usable for fine-grained mechanism claims, which is the honest statement of where the instrument stands.

4.3 The one result that replicates

On the deduplicated main record the composition is deliverable-directed 37 percent, grounds 20, standing 15, frames 15, rule-store 8, channel 5, self-model 1.

No reliability figure attaches to those levels. The kappas above describe blind rater pairs working against the frozen definitions, while the shares are reported under a later scheme version, after recorded adjudication by the author resolved the contested rows, and the corridor that adjudication closed was wide enough to matter: the standing share ran from a fifth to a fourteenth across raters and protocols before adjudication and sits at 15 percent after. The levels are set by the adjudication rather than trimmed by it, so what survives is what holds across the blind runs themselves.

Two results are protocol-invariant across every run. The durability ordering, grounds above standing above frames, holds throughout, with a standing-over-frames margin thin enough that only the ordering and not the gap should be read. And deliverable-directed correction is the largest family.

The second replicates out of sample. The two held-out raters disagreed on how many items were interventions at all, 20 against 30, which propagates into every share and makes composition unestimable at that n. Their profiles diverge on nearly everything. They agree that deliverable-directed correction is the largest family. Nothing else replicates, and the durability ordering does not appear in either held-out profile, though grounds is two items in both, which is an absence of evidence at n=20-30 rather than evidence of absence.

The deliverable result matters beyond its size. It measures a blind spot rather than a behaviour: the field's event ontology types conversational moves against live reasoning, and the largest single category of human oversight labour in this record operates on the produced artifact instead.

5 What the scheme entails: three durability classes

Mechanisms differ in what they supply, and the three classes divide on two questions about the payload: whether a payload exists independently of authority, which grounds and frames answer yes and standing answers no, and whether that payload can be reached by naming it, which grounds answer yes and frames answer no. Absorption follows from those answers rather than defining them.

Grounds-supplying mechanisms revoke or confer entitlement to conclusions. Calling the class informational would mistake the payload for the action: a premise-breaking fact is an undermining defeater (Pollock 1987), and its work is revocation, since one small verifiable fact withdraws entitlement from every conclusion resting on the defeated premise. Demonstration sits awkwardly in the class, since its payload is an exemplar whose criterion resists statement, so the exemplar constitutes the standard rather than reporting one. A constitutive payload has nothing for capability to get better at reaching, and it moves only when a body with standing fixes the criterion by definition and publishes a procedure for realising it, which is how the kilogram stopped being a cylinder. The grounds share is therefore a mixture, one component draining with retrieval and verification against one waiting on a definitional act, and the share's slope over time is a mixture parameter rather than a capability index. What unites the rest of the class is that its grounds live in the world, reachable in principle by better retrieval, wider context, and stronger verification. This class alone is only contingently human, and the contingency is domain-relative rather than capability-relative, measured directly by the corpus, so the contrast that follows carries the corpus's evidence grade rather than a derivation's. 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 class has no self-servable component at all. Absorbability of grounds is a function of the domain's verifier availability, and the phrase takes the verifier as a terminal, a closed check that returns a verdict without deliberating, as a compiler or a measurement does. Verifying a paper, a design, or a diagnosis is a reasoning process instead, and a reasoning process takes grounds, frames, and standing interventions of its own, so stronger verification names a process that may contain the very interventions the class was supposed to shed. That recursion has exactly two terminators, a closed check at some level or an act of standing that declares the verification complete, and peer review terminates in the second way, because an editor decides that review has run its course. Absorbing a grounds move therefore means absorbing its verification stack. A stack that closes in a decidable check absorbs fully with capability, while a stack that closes in standing absorbs down to that terminal act and no further, which is the floor under the compositional prediction of Claim 6. The recursion does not collapse the partition into standing everywhere, since the classes type moves by what they supply while a verification stack is a property of processes.

Frame-supplying mechanisms 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; implicit-choice surfacing converts fixed 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 no reasoner does 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 not absorbable by a single reasoner becoming more capable, and it is partly 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. Different priors have two sources with different fates. Decorrelation is one, and this fraction mechanizes exactly as stated. Consequence exposure is the other: what a framing renders salient may track what its holder can lose. Losses modulate attention (Yechiam and Hochman 2013) and psychological distance changes construal (Trope and Liberman 2010), which makes stake-indexed salience plausible, though neither result shows that a bearer's reframes are unavailable to a non-bearer, and that is the claim that matters. Unlike standing, stake-formed salience is causal history rather than constitution, so it is imitable in principle. Whether it is learnable is what the frame-source experiment of Section 8 exists to settle, and the proposal does not decide it in advance.

Standing-supplying mechanisms are the three zero-content ones, and they differ in kind. 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 a property-rights reading 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 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 purpose re-anchor re-asserts whose purpose governs, which presupposes a purpose-holder distinct from the executor.

The claim can look definitional and is not. 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. Employment was formalized on exactly this object: Simon (1951) models it as the purchase of authority over a zone of acceptance rather than of specified actions.

5.1 The supply condition, and what monoculture actually does

Frames are absorbable because a reasoner with different priors exists to supply them, so the absorber is contingent on a population, not on a capability. The condition is stronger than shared training, because correlation accrues within an episode as well as across a population: two instances of one model in one conversation share not only priors but the frame the conversation has already built, and the narrowing is bilateral, though whether both parties narrow at the same rate is not established here, since a human carries context from outside the episode that the agent does not, which would leave their position partly replenished. A second reasoner helps to the extent it stands outside that loop rather than merely occupying another seat inside it, which makes the decorrelation requirement a claim about position and not about instance count.

Under monoculture that absorber is unavailable, and what follows is narrower than it appears. Monoculture closes the priors route while leaving the procedural ones open, since a rule that redistributes attention and a

change of representation both operate inside a single reasoner's space and neither requires a second population. Frame supply does not vanish; it becomes dependent on invocation. And because a frame cannot be requested by name, the invocation is not a request for the missing content but a decision to run a source and accept what it returns, which makes deciding that a problem warrants the search a threshold the executor does not own. Monoculture therefore relocates the burden onto the standing class rather than restoring frames to the irreducible set, and the durability ordering converges on standing rather than collapsing.

5.2 Not a remainder-humanist claim

The partition risks a familiar misreading, for which Weatherby (2025) supplies the name: remainder humanism, the strategy of locating human distinctiveness in whatever machines cannot yet do, a contest re-run at each capability increment and lost on schedule. The resemblance is structural and the difference is checkable. A remainder-humanist claim indexes its remainder to capability, so it must retreat as capability advances. This proposal cedes the capability-indexed territory in advance: grounds absorb, the decorrelated frame fraction absorbs, and whether the stake-formed fraction follows is open.

Stating that without equivocation requires splitting the standing claim in two. The **predictability** component is empirical: whether a principal's thresholds become learnable from their record. It is capability-indexed, it is what Section 8's designs test, and the proposal stands ready to cede it. The **constitutive** component is positional: a threshold set by the executor is a decision rather than an exercise of the principal's authority, whatever its predictive accuracy, and no experiment addresses it because it is not an effect claim.

The component is not exotic. Peer review instantiates it: an author may not review their own submission, and the rule carries no exemption for authors who could demonstrate calibration, which is what marks it as constitutive rather than epistemic. If the prohibition existed because authors judge their own work poorly, it would be defeasible by evidence that a particular author does not, and it never is. Adjudication states it most strongly, since there the bar survives the disproof of bias: a conviction was quashed because the clerk advising the bench was a partner in the firm prosecuting the related civil claim, on the principle that justice must manifestly and undoubtedly be seen to be done, and the court found no actual impropriety (R v Sussex Justices 1924). Where institutions decline to bar the act, the remedy is still not a better judge but a different one: conflict-of-interest disclosure leaves the assessment where it is and moves the decision to the party whose interest is at stake.

An argument that names both the technical route to its empirical component's failure and the institutional route to its constitutive component's dissolution is not defending a gap. It is pricing one.

5.3 The limit case, and why councils do not settle it

Suppose an agent granted everything the partition names: retrieval exhausting the reachable grounds, a companion reasoner with uncorrelated priors, and an institutional grant of authority with engineered stakes. Custody of any criterion is delegable. Administering a success definition, checking against it, and enforcing its thresholds all transfer. What does not transfer is proprietorship: being the party whose ends make the criterion a criterion. Delivering an interrupt mechanizes while owning the threshold does not. Predicting a preference can exceed the principal's own articulation while its exercise remains performative, since a perfectly predicted consent is not consent.

Mechanism design absorbs administration the same way. Incentive compatibility is defined relative to payoffs that already matter to someone, so a mechanism channels interests and cannot create them, and designating what counts as a payoff is the principal relocated, not eliminated. Chains of standing terminate 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.

That disposes of a tempting substitute. Arrangements in which agents check other agents do real work on the frame class, since each supplies decorrelation to the rest, and none on the standing class, because none of them holds the standing that would have to be conferred. A council of agents is a conduit however many members it has, and adding members lengthens the loss path rather than terminating it. Checks and balances among human institutions work because each branch is separately answerable to somebody outside the arrangement, and the analogy fails precisely where the answerability does. The monopoly case inverts the same point: a single system with no peer has no external bearer to name.

Every external lever of control, liability included, presupposes a party that can be worse off. Stated as a graph condition, deterrence-soundness needs two legs: a loss path from the entity to a bearer with collectability at least the expected harm, and a control path from that bearer back to the entity. Judgment-proofness is the capacity failure of the first (Shavell 1986). An insurer with a loss edge but no control edge fails the second, since premiums get priced while conduct does not change. 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: every agent entity traces to a bearer that both absorbs the loss and holds control.

Adjudication has begun supplying the rule's first leg without waiting for statute. A carrier that argued its customer-facing conversational system was a separate legal entity responsible for its own actions was rejected on exactly that point (Moffatt v. Air Canada 2024). In a second line, liability under fair-housing law was held to reach the supplier of an algorithmic applicant score rather than stopping at the parties who acted on it (Louis v. SafeRent Solutions 2023). Neither is a high appellate holding, one being a small-claims tribunal and the other an interlocutory ruling later settled without admission, so the weight is directional rather than settled. The direction is the one the graph condition predicts: courts decline the move that severs the loss path, whether the severing device is a claim of separate entity status or of intermediate causation.

6 A formal statement, and the assumption it rests on

Under incomplete contracting, expected loss from a misset delegation threshold converges to a strictly positive floor, with the standing class exposed as conditional on non-enumerability that persists under learning. The compositional prediction of Claim 6 falls out as a corollary. The full statement and its proof are in Paper B; what matters here is the antecedent, because the antecedent is where the argument is vulnerable.

The model turns on an information set. Capability improves inference from the information held, and retrieval or wider context enlarge that information, which is what makes the grounds class absorbable. But enlarging an information set is learning within a fixed state space, so the argument presupposes that the space is given and shared. That is the small-world assumption, and the frame class is precisely its violation: a frame replaces the hypothesis space rather than the information indexed over it.

That is the small-world assumption Savage's framework was criticized for, and which Karni and Viero (2013) identify as the reason a choice-theoretic treatment of unawareness needs a more general point of departure. Decision theory formalizes the move as growing awareness, and the distinction it draws is this proposal's own, since the discovery of new acts and consequences expands the universe while new evidence about links updates within it. The formal result there is that relative likelihoods of events in the original space survive the expansion, which supplies the frame class with apparatus this argument does not itself provide. The two are complementary, and the boundary is where each is silent: that literature models what follows once awareness has grown and takes the discovery as given, treating state spaces as exogenous. This proposal asks the other question, what supplies the expansion and whether the supplier is absorbable.

Answering it requires separating the sources, because they differ in what they change. Different priors are one, the only route that alters what is reachable in principle. A procedure is a second, redistributing attention within a space the reasoner already spans; considering the opposite is the canonical instance, and its measured advantage runs over generic instructions to be fair and unbiased (Lord, Lepper and Preston 1984), which locates the operative ingredient in the structure rather than the exhortation. Procedures mechanize completely and cannot surface what the space does not contain. Re-representation is a third, moving possibilities across the boundary of practical rather than principled reachability. The fourth supplies no content and gates the other three, being whatever causes any of them to run on this problem rather than in general.

To these the unawareness setting adds a constraint. A frame cannot be requested by name, since a request specifies what it seeks and naming the missing possibility would require already representing it. Expansion is solicited only obliquely, by invoking a source and accepting what it returns, which makes the fourth source operative for the class as a whole and locates it outside the frame class entirely: deciding that a problem warrants the search is the setting of a threshold, which Section 5 places in the standing class.

That routing does not collapse the partition, because the classes divide on the payload rather than on the trigger. Deciding that a claim warrants a lookup is threshold-setting on the same argument, and the move the decision triggers is still grounds-supplying. A frame's payload is apt or inapt independently of whoever decided to seek it, where standing's residue is an act with no fact behind it to reach.

7 Why the existing instruments cannot settle this

The compositional prediction of Claim 6 is a claim about what aggregate measures will show, so the natural first move is to test it against the best aggregate measure available. That move fails, and the reasons are informative enough to become part of the case for the program.

The Anthropic Economic Index is the relevant instrument: a recurring open dataset mapping real conversations to occupations and tasks, with seven releases between February 2025 and June 2026 (Handa et al. 2025, Anthropic 2026). It classifies each conversation into one of five interaction types, grouped into automation and augmentation. From the March 2026 report: "Since our first report, we have classified conversations into one of five interaction types—directive, feedback loop, task iteration, validation, and learning, which we group into two broader categories: automation and augmentation."

The axes are orthogonal, not competing. The Index partitions by division of labour and direction of flow; this proposal partitions by what a human message supplies. Sorting the five categories by what flows and in which direction makes the consequence visible. Two of them, validation and learning, describe AI-to-human transfer and contain no human interventions in this proposal's sense. A third, feedback loop,

describes the agent consuming environmental feedback, which is the canonical case of grounds supply absorbed by tooling rather than supplied by a person. A fourth, directive, contains human interventions only at its endpoints. That leaves task iteration as the single cell holding this proposal's object in quantity, and inside it the correcting fact, the reframe, and the stop are all collaborative refinement. The mapping is many-to-many with nothing isomorphic in either direction, and the field's largest augmentation category is a sum over classes with opposite absorption trajectories, so its share can hold steady while its composition inverts.

The instrument is least sensitive to the class this argument says is durable. The automation-augmentation ratio tracks conversational volume. Grounds supply is verbose: supplying a missing fact, pointing at evidence, correcting a premise are each messages with content, and a run of them reads as collaborative refinement. Standing supply is the opposite. A stop, a judgment demand, a purpose re-anchor carry no domain content and consume almost no exchange, and accepting a deliverable consumes none. So as grounds absorb, conversations lose exactly the messages that made them look augmentative and migrate toward directive, while the residual human contribution, being zero-content, registers as near-silence to an instrument that infers involvement from exchange. The Index would record a clean rise in automation share across precisely the period in which threshold-setting was unchanged.

The measure is not wrong on its own terms. It measures division of labour, and by that measure the labour did shift. But an index whose sensitivity to a class runs inversely to that class's verbosity is structurally least able to detect the one class this argument says does not absorb. That is Claim 6 derived from the Index's own published definitions rather than asserted against it, and it needs none of the Index's data.

The longitudinal test is not available, and the reasons are the specification for what to build. The obvious check is whether task shares hold steady while mode composition moves across the release series. Three confounds block it. The task axis is broken by a coding vintage change, since the March 2026 report uses 2019 O*NET-SOC codes where previous reports use the 2010 vintage, and that crosswalk is many-to-many. The corpus composition changed when Claude Code data entered the pool, and agentic sessions carry a different interaction profile from consumer chat. Most seriously, the mode axis is uninterpretable because the classifier is undisclosed: the taxonomy is stable across releases, which is the one favourable finding, but the report does not say which model performs the interaction-type classification or whether it was held constant, and that classifier is itself a model improving over the same window in which the effect of improving models is the quantity of interest. Any trend is contaminated by drift that co-moves with the treatment.

A null would not have been informative and a positive would not have been credible, so the test was designed, gated, and withdrawn rather than run. What survives is a specification. Settling the compositional question needs a message-granular instrument with a disclosed and version-pinned classifier, applied to a corpus whose composition is held constant or measured. No existing dataset supplies all three, which is the first thing Section 8 proposes to build.

8 The research program

Six lines, ordered by what each would settle and what it costs. The first three are runnable now on public data; the next two need a harness at the scale of a funded program; the last is a long-horizon design worth pre-registering early for a reason stated below.

8.1 Validation of the scheme, on public corpora

The binding limitation of the pilot is that its corpus is one practitioner's, and the out-of-sample check exhausted the available held-out material at n=93. External validity needs a second corpus at message granularity from other people's transcripts.

Re-code public interaction corpora with the frozen definitions, measuring coverage, inter-rater agreement, and the frequency and outcome of each mechanism. This requires no private data. The coding pipeline has published precedent: privacy-preserving taxonomy extraction over conversation corpora at scale is demonstrated infrastructure (Tamkin et al. 2024), already applied to build a hierarchical values taxonomy from three hundred thousand conversations.

Two design rules carry over from the pilot's failures. Verification must be acyclic, with raters never seeing the intervener's reasoning, or the check inherits the author's framing. And the pilot's family agreement of 0.58 out of sample sets the bar this study has to clear before any mechanism-level claim is reportable; the gate figure of 0.82 is what licenses the compositional work in the meantime.

8.2 The falsifier: the three-arm interrupt experiment

Claim 7 names the condition under which the standing class is misidentified, and the design that tests it separates the candidate readings of why zero-content moves work. Identical contentless interrupts are 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, and the

partition is wrong.

– 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.

– The script arm splits by schedule. Interrupts timed by a policy learned from the

principal's intervention history test the predictability component directly. Interrupts on a random schedule test whether even the timing carries intelligence, which is the correlation-device reading in pure form.

– Only **principal-arm superiority** preserves the strong authority reading, and the

partition predicts it.

The empty cell is stated in advance, with the scope its harness imposes. If no arm moves outcomes on tasks whose success criteria are fixed beforehand, the zero-content moves were not causal inputs to success in the region the decomposition already expects to absorb them. That weakens the forcing-function reading without refuting it and leaves the constitutive component untouched, since a move whose function is authorization was never predicted to improve search. A refutation of the forcing-function reading requires a null where criteria are not enumerated in advance.

A control-condition pilot already prices this study rather than discouraging it. Eighteen tasks across two families, self-contained synthesis and debug-from-spec with defects planted to survive their own visible tests, fifty-four runs, fifty-three passing with one run-to-run disagreement in total. Two design facts follow: repeated runs of a fixed task carry almost no information at this scale, so sample size must come from task count rather than repetition, and difficulty in this family did not rise with defect subtlety. A discriminating test needs long-horizon tasks in real repositories with requirements not fully specified in advance, and

outcome scoring by judges under an acyclic protocol.

8.3 What the frame class's absorber actually is

Claim 3 turns on whether a second reasoner supplies frames because it holds different priors, or merely because it stands outside the accumulated loop. Two experiments separate the possibilities.

A **stance-versus-source** experiment compares reframes produced by one model asked to occupy an outside stance, by a fresh instance of that model with no shared episode history, and by a model of different lineage. That separates the three things the absorber could turn out to be: a prompted posture, an escape from accumulated loop state, or genuinely different priors. It is decisive for the monoculture condition in either direction.

A **frame-source** experiment extends the same logic to content-bearing interventions, comparing reframes supplied by a stake-holding human, a disinterested human, and a second model. Stake-indexed salience predicts divergence on distributional rather than technical frames. One caveat is structural: stake-formation is source-identified rather than transcript-codable, so the split is testable experimentally and not observationally.

8.4 The composition index

Claim 6 is the policy-facing claim and Section 7 establishes that no existing instrument can test it. The measurement is an oversight-composition index over interaction corpora, tracking the grounds, frame, and standing fractions across model generations.

The estimable specification is 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. The design is conservative under the most likely selection story, since assigning harder tasks to better models biases the grounds coefficient toward zero.

Three requirements follow directly from Section 7's failed test: a disclosed and version-pinned classifier, a corpus whose composition is held constant or measured across periods, and message granularity rather than conversation aggregates. A training-time flank needs handling too, since a model fine-tuned on a practice's own memory store migrates enumerated grounds into weights, so the index should treat store-trained deployments as a generation increment and predicts the same signature.

8.5 The harness, which is worth building beyond this paper

Sections 8.2 and 8.4 both require infrastructure that does not exist: long-horizon tasks in real repositories with under-specified requirements, judge-scored outcomes under an acyclic protocol, and instrumented interventions typed at message granularity.

That harness is the general instrument for measuring what oversight contributes, and no existing benchmark supplies it. Its value does not depend on this framework being right. Any account of human-AI collaboration that makes claims about what human input contributes needs a way to vary that input and observe the consequence, and the field currently evaluates agents on tasks where the human contribution has been designed out in advance.

8.6 The cohort study, and why it should be pre-registered now

If the composition inverts, the consequence for oversight competence is a workforce question with a long lag. The mechanism is generational rather than individual. Procedural work absorbed by agents is also the consequence exposure on which the next cohort's judgment 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. Entry-level task absorption can be dated, which makes a difference-in-differences design over occupation, entry cohort, and period feasible, 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. In software, oversight competence is proxiable by defect-catch rates in review, review depth conditional on diff size, and time to review authority. Whether a formed operator's own intervention skill decays under a deferent partner is a separate claim, unmeasured, held as a hypothesis and outside this design.

The binding constraints are cohort selection, proxy validity, and a five-to-ten-year lag that leaves the design underpowered until roughly 2030. That lag is the argument for pre-registering now rather than later: the pre-absorption cohorts are the control group, and they are no longer being replenished.

8.7 Two lines held as conjectures

Stated as conjectures rather than claims, because their mappings are not yet defended. Read onto this setting, the delegation model of Aghion and Tirole (1997) 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 that model's assumptions to a human-agent loop must be constructed and defended before the test carries weight. Until then it is a candidate cost model, not a result.

Separately, the maintenance stream created by the migration of interventions into standing structures is the quantity of record for the Baumol restatement in Section 9, and a time-and-cost accounting of it is the measurement that would price it.

9 Limitations

The corpus is one dyad. One practitioner, one agent lineage, retrospectively coded by the collaboration under study. Retrospectives over-sample correction-rich sessions, so nothing here estimates intervention frequency or the base rate of sessions needing none. The transcript-completeness failure of Section 4.1 applies to the corpus itself: conversational mechanisms are better sampled than out-of-band ones, and cross-mechanism frequency comparisons are unreliable.

Two assumptions are load-bearing and unexamined. The scheme assumes interventions are received and honored. It classifies what an input supplies and asks what can absorb it, and does not ask what happens when an agent capable of evading a restriction declines to comply. That case needs a threat model this proposal does not develop and is scoped out rather than treated thinly. The scheme also assumes the intervener is right. The corpus contains corrections that over-generalize, and instances in both directions where the agent refused or refuted a human premise and the human accepted it, so resistance to a wrong intervention occurs. What is missing is a protocol that makes it reliable rather than incidental, and model deference is what makes the gap dangerous rather than merely open, since a wrong intervention looks identical to a right one from the intervener's side.

The event ontology excludes part of its own subject. The scheme 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. A maturing practice moves the subject matter out of the channel the scheme observes, which yields a confound worth naming, since conversational intervention counts should fall as a practice matures at constant capability. The rule-store family, at 8 percent of classed interventions, is the conversational trace of a stock that grows as the per-session counts fall.

The scheme has no term for the verification a move rests on. It types moves by what they supply, while the recursion of Section 5 makes residue a property of processes as well as of moves, and a stack that closes in an act of standing carries oversight labour the move's own class never records. Composition totals summed over coded moves therefore undercount standing wherever verification terminates in it, and further coding does not close the gap, since the missing quantity is not a move.

The magnitude question has a structural frame and no measurement. When one component of a production process resists productivity growth while its complements accelerate, the resistant component's cost share rises even as its volume falls (Baumol 1967). The composition inversion is that result at message granularity, and it relocates the de minimis objection, since the question is never whether standing labour is currently small but whether it is the component whose share the schedule inflates. The corpus supplies a count-based anchor and nothing more: standing runs from roughly a fifth to roughly a fourteenth of coded interventions across raters and protocols before adjudication, and sits at 15 percent after, with time shares unknown and likely lower because zero-content messages are short. A naive application would also contradict this proposal's own migration mechanism, since standing labour that migrates into set-once owned parameters exhibits productivity growth rather than stagnation. The Baumol candidate is the maintenance stream the migration creates, not the per-episode message stream.

The literature survey is access-filtered and structurally biased. Paywalled sources were undersampled, retrieval skewed toward self-archived preprints, and coverage is English-language throughout. The structural bias is worse: the survey searched on this proposal's own coinages while arguing that the field lacks the vocabulary, so any field naming 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, psychotherapy process research, aviation crew resource management, and legal practice, the last being a probable prior home for the judgment demand specifically. Every novelty claim should be read as scoped to the surveyed literatures. One reconciliation step exists: one unmapped class, the repair of misresolved referents, is near-verbatim other-initiated repair (Schegloff, Jefferson and Sacks 1977), which converts one member of the threat list into an import and leaves the rest standing.

The reflexivity is disclosed rather than resolved. This document is agent-assisted, inside the same collaboration the corpus documents, with claims, codings, and revisions directed and reviewed by the author.

References

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

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

Baumol, W. J. (1967). Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review* 57(3).

Buterin, V. (2025). AI as the engine, humans as the steering wheel. https://vitalik.eth.limo/general/2025/02/28/aihumans.html

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

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).

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).

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

Karni, E., and Vierø, M.-L. (2013). "Reverse Bayesianism": a choice-based theory of growing awareness. *American Economic Review* 103(7). https://doi.org/10.1257/aer.103.7.2790

Lord, C. G., Lepper, M. R., and Preston, E. (1984). Considering the opposite: a corrective strategy for social judgment. *Journal of Personality and Social Psychology* 47(6). https://doi.org/10.1037/0022-3514.47.6.1231

Louis v. SafeRent Solutions (2023). No. 1:22-cv-10800 (D. Mass.), order largely denying the motion to dismiss; settled 2024 without admission of liability.

Moffatt v. Air Canada (2024). 2024 BCCRT 149 (British Columbia Civil Resolution Tribunal).

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

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

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

R v Sussex Justices, ex parte McCarthy (1924). [1924] 1 KB 256 (King's Bench Division).

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

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

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

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

Trope, Y., and Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review* 117(2). https://doi.org/10.1037/a0018963

Weatherby, L. (2025). *Language Machines: Cultural AI and the End of Remainder Humanism*. University of Minnesota Press.

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

Yechiam, E., and Hochman, G. (2013). Losses as modulators of attention: Review and analysis of the unique effects of losses over gains. *Psychological Bulletin* 139(2). https://doi.org/10.1037/a0029383