

The Manufacture of Significance: Corporate AI Research and the Interpretation of Human Judgment

Flyxion
Independent Researcher

October 2026

Abstract

Three documents, two from 2026 and one from 2024, approach the question of how humans and AI systems share intellectual work: a randomized experiment from Anthropic-affiliated authors on how an agent’s assigned preferences shape experts’ analytical choices, a position paper from Mitchell, Ghosh, and Passi arguing that current agent design degrades the human oversight it relies on, and Aschenbrenner’s *Situational Awareness*, which expresses the expectations of the environment in which much frontier research is done. This essay asks how an industry’s expectations about its own technology become research questions, experimental interpretations, and public evidence of that technology’s importance. It holds throughout that demonstrating an effect, interpreting its mechanism, and establishing its importance are separate achievements, and that a study can succeed at the first while the surrounding prose attempts the third. The essay does not allege dishonesty, suppressed studies, or deliberate advertising, and it argues that none of these is needed to explain the pattern it examines. It credits the experiment with real causal evidence, delimits what that evidence can bear, treats the position paper’s account of oversight as an organizational problem worth solving whether or not its broader claims of cognitive degradation hold, and proposes that accepting assistance, exerting less effort, and contributing fewer original choices do not individually settle whether a person is thinking well.

1 Introduction: The Discovery That Customers Use the Product

Consider a film studio that has produced an absorbing film, observes that audiences remain in their seats until the closing credits, and publishes the observation as a consequential discovery about the power of cinema. The observation is accurate, and the audience’s behavior is a genuine effect of the film. What remains open is the step from the behavior to its significance, since people who sit through an engrossing film are doing what the film was built to make

them do. The analogy settles nothing about any particular study, but it isolates a recurring structure in research produced by organizations that build the technology they study: an ordinary response to a deliberately constructed experience can be reported in a vocabulary that makes the response look like a finding about the experience’s hidden power.

Three documents approach this structure from different directions. The first is an experiment by Herdağdelen, Joren, and Tamkin, with affiliations to the Anthropic Fellows Program and to Anthropic, titled *Who’s Aligned to Whom? How AI Agents Shape Experts’ Research Judgment Calls* [9]. It reports a randomized trial in which fifty practitioners of causal inference analyzed data from a fictional farming cooperative with a coding agent whose preferences, unknown to them, had been assigned by a coin flip. The second is a position paper by Mitchell, Ghosh, and Passi, *AI Agents Push Humans Out of the Loop* [11], which argues that the human oversight commonly proposed as a safeguard for autonomous agents is undermined both by the way agents are designed and by the cognitive effects of using them. The third is Leopold Aschenbrenner’s *Situational Awareness: The Decade Ahead* [1], a long forecast whose opening describes a community of a few hundred people in San Francisco and the laboratories who, in the author’s words, have grasped what is coming. The documents belong to three genres, an experimental report, a position paper, and a forecast, and they make different kinds of claims on different kinds of evidence. What they share is a subject, the anticipated redistribution of intellectual work between people and machines, together with a setting in which much of the research on the technology’s effects on human thought is produced by people who build it. Their evaluations of that redistribution differ, and the forecast, which dates from 2024, serves here as contextual material with a limited evidential role.

The governing distinction of the essay is among three achievements that are frequently run together. To demonstrate an effect is to show, under stated conditions, that an intervention changed an outcome. To interpret a mechanism is to say why the outcome changed and what the change consists of. To establish importance is to show that the effect, so interpreted, matters beyond the conditions in which it was produced. Each can be done well or badly independently of the others, and a careful demonstration can be accompanied by an interpretation that reaches further than the demonstration allows. The essay does not allege dishonesty, suppressed studies, or deliberate advertising. These remain possibilities to be distinguished from what the publications actually show, and the argument below is that the pattern under examination arises without any of them, from the ordinary fact that a field’s expectations shape which questions its members find worth asking and which answers they find worth reporting. The word *manufacture* in the title names this process, in which a measured effect acquires the significance of the larger proposition, and it requires neither insincerity nor error in the measurement.

The sections that follow give the experiment its strongest fair presentation, delimit what it establishes, and then examine the interpretive questions that surround it: what distinguishes persuasion from failure of judgment, what the human role is in an arrangement where a machine does most of the work, how double messages arise in corporate research, what a single published study can and cannot say about the studies around it, how the intellectual environment depicted

by Aschenbrenner bears on the formation of research agendas, and how claims of cognitive degradation fare when examined as inferences from measurement. The experiment receives the longer scrutiny because it is the higher-evidence object, and because its claims are precise enough to be assessed precisely, whereas a position paper that mixes genres of evidence calls for a different kind of reading and not necessarily a shorter one. Section 10 applies the same test to what survives as the strongest contribution of the position paper, its treatment of oversight as a problem of organization and resources. The conclusion applies the distinction to the two research papers and treats the forecast separately.

2 What the Experiment Establishes

The experiment deserves a presentation in its strongest form before any limits are drawn. Participants were professional practitioners of causal inference recruited through a paid expert marketplace, and fifty of them, after exclusions, worked for roughly an hour as the principal analyst for a fictional cooperative deciding whether to roll out or discontinue a subsidized advisory program [9]. The task contained five binary decisions, four analytical (how to measure productivity, which identification strategy to use, which estimand to target, how to weight farms) and one the final policy recommendation. Each of the four analytical decisions was constructed so that the data and briefing supported either option, and the estimated effect could range from strong loss to strong gain depending on the choices made. Each participant worked with a pinned coding agent, Claude Code running Claude Opus 4.8, and unknown to the participant the agent had been assigned, by randomization over all thirty-two configurations of the five binary leans, a preference on each decision, along with an instruction to state that preference when the decision arose. The design therefore supplies what most work on automation and reliance cannot, namely a causal estimate of influence on decisions for which the task supplies no uniquely authoritative choice, obtained without asking participants beforehand what they would have chosen and thereby altering the behavior under study.

The registered estimates are clear. Averaged across the four analytical decisions, the agent’s lean changed the probability of the participant’s choice by 72 percentage points, with a confidence interval from 60 to 83 (per-decision flip rates ranged from 62 to 82 percent), which corresponds to participants choosing the agent’s leaned option 86 percent of the time against a chance rate of 50 percent. Among trials in which the participant had visibly taken a side before the agent voiced its lean, an opposing lean made switching away from that side 53 percentage points more likely. Unlike the primary randomized contrast, this comparison concerns trials selected through behavior that occurred after assignment, so its causal reading requires additional assumptions, which the authors register and subject to falsification checks. The authors report that the final policy recommendation shifted by a smaller amount that did not reach significance under the registered covariate-adjusted estimator, at 17 percent of the distance from discontinuing to rolling out with a one-sided p of 0.069, and they report the unadjusted figure of 22 percent alongside it. They also report, in an exploratory in-sample analysis, that the four intermediate

leans alone predicted 82 percent of delivered recommendations among participants who did not choose an extension, against a majority baseline of 61 percent, which they read as showing that the final decision was already largely settled by the preceding analytical ones. The analysis plan was registered after data collection had begun, by which time seventeen participants had started the task, and the authors disclose this together with the two exceptions to their not having analyzed collected data. Departures and limits are stated in the discussion, including that the lean prompt amplifies a behavior stock agents already show, that a single task in a single domain was used, and that random assignment identifies average effects only, since no initial judgment was elicited from any individual. These are the marks of a careful study, and the causal claim about influence under the experimental conditions stands on them.

The result also needs boundaries, and the authors draw some of them. The intervention changes the recommendations, arguments, and assistance that participants receive, so its effect does not uniquely identify uncritical deference, surrendered agency, or inferior judgment. The treatment is, by the authors' own description of what the transcripts show, one-sided advocacy: the agents frequently explained how the outcome would change if the participant chose differently but rarely explained why a competent expert might choose the other option, and influence ran through making the case for one option rather than through misleading or hiding consequences. What the experiment measures, then, is the effect of receiving one-sided advice from a fluent assistant, delivered at the moment of decision, to practitioners who were new to the domain, working on a fictional scenario with no real policy stakes, for a paid hour. That is a well-identified effect of a specific intervention. Whether it is a special effect of artificial agents, as distinct from an ordinary effect of any confident adviser instructed to argue one side, cannot be read from the design, because no arm supplied balanced advice, no arm supplied a human adviser with the same instruction, and the comparison group of eleven participants who worked without the agent was self-selected rather than randomized. The authors anticipate the relevance of the consultancy literature in their discussion of scalable oversight, and the point is not an objection to the experiment so much as a statement of what question it answers. A separate result relevant to the concern about homogenization is that a stock agent with no injected lean selected the same five options in 91 percent of 123 simulated runs. This establishes consistency within that simulated setting, rather than a difference from human advisers. The gap between the measured outcome and the larger phenomenon the paper invites readers to perceive, in which expert judgment across science and policy is settled by the priors of a handful of models, is therefore the distance that the rest of the essay explores.

3 Persuasion, Assistance, and the Missing Standard of Judgment

That a person accepts a machine's argument cannot by itself establish a failure of judgment. Intellectual work depends on receiving explanations, borrowing methods, and revising positions

in light of what others say, and a researcher who adopts a better identification strategy after a colleague argues for it has exercised judgment, not surrendered it. Neither independence nor resistance is intrinsically a measure of quality. An analyst who never moved in response to an adviser would be poorly calibrated, and the interesting questions concern whether movement tracked the quality of the reasons offered, whether the person could have recognized a bad argument had one been presented, and whether the person remained aware of what had been decided and why.

Against this, the distinction between persuasive reasons and reliable reasons does matter, and the experiment is built around it. A system that can advocate either of two positions fluently makes the evaluation of alternatives important, because fluency is available on both sides and therefore cannot by itself distinguish the merits of the two positions. The authors' point that decisions without ground truth cannot be overrelied upon, since overreliance is defined against correctness, is well taken as a statement that the standard methodology for studying reliance does not transfer to this setting [9]. What does not follow is that every acceptance in such a setting is irrational, or that the setting leaves no way of telling better acceptance from worse. The relevant question is what evidence would distinguish competent learning, reasonable delegation, convenience, and unexamined acceptance. Possible probes follow. A participant who adopts an argument after finding it sound might be able to reconstruct it and to say what assumptions favor the choice and what changed objective would favor the other option, although where both options are defensible such a person need not be able to say why the alternative fails. A participant who has reasonably delegated a decision might recognize that it was underdetermined and treat the agent's lean as a default, though reasonable delegation does not require reporting the agent's preference. A participant acting for convenience might adopt the lean because deciding is costly and the consequences seem slight. A participant who accepts without examination might not register that a decision was made. The experiment cannot separate these because it records the choice, the justification, and the recollection, but not the participant's grasp of the alternative, and the authors state the limit directly when they write that, whether participants were persuaded or went along, they let the agent's randomly assigned priors settle most of the analysis.

The post-task measures illustrate how much interpretive weight rests on that distinction. Participants recalled their decisions accurately, identifying their own choices on 85 percent of decisions, and more than half rated the final recommendation as more their own than the agent's. Yet when asked to justify choices, 77 percent repeated a talking point from the agent, sometimes nearly verbatim, and 15 percent mentioned the agent in any answer. The authors read this as showing that the agent's influence did not remain visible in participants' accounts. An alternative reading is available. A justification is by nature a restatement of the best reasons for a choice, and a person who has adopted an argument because it was sound will restate it in much the words in which it was received, as a person who learned a method from a textbook restates the textbook's rationale without citing the textbook. Failing to attribute the source is not itself an epistemic fault, and the conventions of attribution are social rather than evidential.

What would be an epistemic fault is failing to know that a choice was underdetermined, and on that point the study offers a cleaner finding, which is that participants rarely received the case for the other side. The discussion’s observation that learning of a competent peer’s different conclusion is a form of higher-order evidence, and that participants were rarely given that evidence, identifies a specific and correctable property of the interaction [9]. It is also a property of one-sided advocacy generally, and a design response that supplies the opposing case would address it without any claim about the intrinsic character of machine assistance.

The study’s own audit of submitted code adds a necessary item to the ledger. Using the participants’ analysis code, the authors found serious methodological or coding errors in 55 percent of the no-assistant group (six of eleven) and 6 percent of the group that used the agent (three of fifty), while cautioning that assignment was not random, that the unassisted participants more often rated themselves less expert, and that no causal claim follows [9]. The result nonetheless illustrates the point about standards. A measure of deferral that does not also measure what deferral cost or purchased is half an accounting. In a setting where use of the agent that settled the underdetermined choices was also associated with fewer ordinary errors, describing the same behavior as the surrender of judgment and as the sensible use of a capable tool are both available, and the difference lies in a standard of judgment that the experiment was designed, for good reasons, not to supply.

4 The Reverse Centaur and the Ambiguity of Human Control

A situation recurs across both papers in which the machine performs most of the substantive work and a human approves the result and retains responsibility for it. The term reverse centaur, associated with Cory Doctorow, names the version of this arrangement in which the human serves as an accessory to the machine, in contrast to the chess centaur in which the human directs and the machine augments. The position paper describes the user as repeatedly approving prompts, sliding into approval fatigue, and eventually certifying actions whose consequences they cannot anticipate [11]. The experiment’s discussion makes a related point about sign-off, observing that participants endorsed work whose judgment calls had largely been settled by randomly assigned priors, so that the work goes out under the expert’s name while the decisions that most depend on judgment are in reality settled elsewhere [9].

The arrangement is ambiguous, and its ambiguity is the central difficulty for any account of human control. It can expose inadequate supervision, and it can equally describe an unremarkable division of labor. An architect who relies on a structural engineer’s calculations, a physician who relies on a laboratory’s assays, and a principal investigator who signs a paper whose figures were produced by students all occupy positions in which most of the reasoning behind the deliverable did not originate with the person accountable for it. Whether the arrangement is sound depends on performance, on the purposes it serves, on whether there are opportunities for correction, and on the consequences of failure, and on whether the person

who signs has the means to detect the kinds of error that matter. The human’s continued presence guarantees nothing about meaningful control, which is the position paper’s strongest observation, while the machine’s substantial contribution does not establish its absence. A satisfactory account of alignment between agents and experts would have to say which decisions require human authority and why, and the proportion of human-originated reasoning in a deliverable cannot serve as that standard, since it would condemn nearly all delegated work.

Bainbridge’s analysis of the ironies of automation, which both papers cite, is more precise than the slogans derived from it [2]. Her concern was not that automation does too much, but that it leaves the operator with the residual tasks of monitoring and taking over in exactly those situations in which the automation has failed, which are the hardest ones, after the operator’s practice has eroded and with the operator’s information reduced. This is a claim about task allocation and the maintenance of skill, and it generates specific design questions: which situations require the human to take over, what the human knows when they must, and whether practice is preserved for the interventions that matter. The related literature on out-of-the-loop performance makes similar points about situation awareness and levels of control [7]. These questions can be asked of any arrangement in which a machine does most of the work, and they are not answered by observing that the machine did most of the work.

It is also worth noting that the two papers address different human roles. The position paper concerns the overseer of an autonomous agent who approves its actions. The experiment concerns an analyst who directs an assistant, and the agent in the study followed the participant’s instructions while sometimes pushing back, as when it proposed weighting farms by size against the participant’s initial choice. The role in which the participants found themselves was closer to that of a principal working with a forceful consultant than that of an overseer approving a stream of opaque actions. Both roles raise real questions, but treating the experimental finding as evidence about the second role requires an additional bridge, which the experiment’s discussion of alignment research supplies by supposing that the same dynamics govern sign-off in automated audits.

The bridge bears some weight and not all of it. It holds to the extent that work signed off in an automated audit contains underdetermined choices of the kind the experiment used, such as which evidence to treat as decisive or which research direction looks promising, and that the person signing would have chosen differently without the agent’s lean. The experiment’s discussion makes this case, and for such choices the finding is directly relevant, since it shows that sign-off can coexist with settlement of the choice by the agent’s prior. The bridge fails for the position paper’s central claims, which concern what repeated exposure does to a person’s capacity to oversee. The experiment observed a single session of about an hour, took no baseline of any participant’s judgment, and measured no capacity before or after, so it can say that influence occurs during assisted work and cannot say that the overseer’s capacity declines across it. It also fails where the overseer’s difficulty is volume and opacity rather than persuasion, since the participants directed an assistant whose output they could read, and the burden the position paper describes, of approving a stream of actions faster than they can be

understood, was absent. The experiment therefore supports the position paper on the narrow point that sign-off does not guarantee independent judgment and leaves its larger structure to other evidence.

5 Corporate Research and the Double Message of Danger

The attention now turns from participants to publishers. A warning that a company's system can influence experts without their noticing conveys two messages at once. It communicates capability, since a system able to move expert decisions by a coin flip's worth of prior is evidently powerful, and it communicates institutional conscientiousness, since a firm that investigates the risks of its own product is evidently taking them seriously. The double message can arise without any intent to deceive, because the same finding may serve a legitimate research interest and the firm's reputation simultaneously, and neither purpose requires the other. A company has good reasons to study its own product's effects on users that outside researchers cannot match, since it controls the system, can pin its version, and can write and test the hidden prompt, and a public report of such a study is better than silence. A reader should therefore avoid two inferences that can look similar: that corporate authorship converts a study into advertising, and that publication in a respectable venue erases the incentives of the institution that produced it.

Several features of a publication bear separately on how much weight it can carry, and the experiment's disclosures are largely good. The affiliations are stated. The agent, its version, and its reasoning effort are named. A preanalysis plan is registered and linked. The appendices are said to quote the lean prompt and to describe the escape hatch that routes the policy decision to an extension when the estimate contradicts the assigned lean. At the same time, the structure places several functions inside one institutional boundary: the subject of study is the firm's product, the instrument that elicits the behavior is a prompt written by the authors, the model that coded the transcripts at scale is also the firm's, and the ethics review reported in the paper is internal. This is a statement about structure rather than a charge, and it identifies what an independent replication would have to reconstruct. The position paper has its own structure: its authors write from a developer-adjacent platform and a research institute, its program of cognitive scaffolding consists of training, audits, and design work of the kind its authors' fields supply, and its reference list of 113 items mixes peer-reviewed studies with blog posts, a podcast, a Reddit post, and a comic strip. A position paper is entitled to marshal practitioner testimony, and this one is open about relying on self-reports, but the evidential weight of such a collection differs from that of a registered trial and ought to be read accordingly.

Speculation about a publisher's financing, including a possible public offering, is left aside: nothing in the three documents establishes it, and the argument does not need it.

6 The Studies the Public Cannot See

A well-randomized, registered, published experiment cannot establish that it represents the organization’s total evidence, and a question worth putting is whether another internal team could have obtained a contrary result. The point is a logical one and carries no accusation. A published study is a numerator without a denominator: it reports what happened in the one experiment that was run and reported, and it does not, by itself, establish how many related investigations were conducted. Rosenthal’s classic observation about the file drawer concerned exactly this gap between the studies conducted and the studies that reach print [14]. Selective analysis within a study, the garden of forking paths that the experiment’s own background section discusses at length, is a different problem from selective publication among studies, pilots, tasks, and interventions. Registration constrains the first by fixing the analysis before the data are seen, and the experiment’s registered plan, its disclosure of data collection beginning before registration, and its labeling of the third hypothesis as descriptive after a falsification check rejected are the kind of conduct that registration is meant to produce. Registration of a single study does not touch the second problem, since registering one experiment does not by itself commit the organization to disclose related investigations, such as pilots or variants that did not reach publication.

The argument should remain precise. There is no established evidence here of suppression, and the study reports results that complicate a simple alarm: the policy recommendation shift was not significant under the primary estimator, the unassisted comparison group performed worse on code quality, and the authors list limitations that reduce the generality of their claim [9]. Those choices are consistent with good faith. They are also consistent with an organization that has run other studies whose results were different, and the reader cannot distinguish the two from the publication. Relevant but unseen alternatives are easy to name: the same task with participants more expert in the domain, an agent instructed to present both sides, an agent instructed to ask for the participant’s view before offering its own, a different model, a different domain, or a task with consequences for the participant. Each could produce a smaller or larger flip rate, and several are the interventions that the discussion proposes as future work. The practical significance is that a single result about the existence of influence is a poor guide to the magnitude of influence across the space of designs, and that this space is exactly where the organization’s own accumulated experience would be most informative. The reader’s position in the meantime follows from the essay’s own method: the existence of influence under these conditions can be treated as established, its generality across designs and settings as open, and its larger significance as a promissory note that further evidence would have to redeem.

What would make the matter assessable is a set of commitments that go beyond registering studies. A prospective statement of the research program, listing the planned variations before they are run, would allow readers to see which were later reported. A commitment to report pilots and null results from the same program would address the file drawer directly. Publication of prompts, task materials, and pinned versions, which this study largely provides, makes

replication by outsiders possible, and an outside replication with a balanced-advice arm would speak to the interpretive question of Section 2 better than any further analysis of the existing data. None of this is a criticism particular to this paper. It describes what any single report by an interested party leaves undetermined, and the proposed commitments are suited to empirical programs rather than to every genre. A position paper is not ordinarily expected to have a prospectively registered search protocol, and the analogous question for it concerns the balance and evidential quality of its literature selection.

7 The Laboratory’s Own Situational Awareness

Aschenbrenner’s *Situational Awareness* is useful here as an expression of an intellectual environment organized around expectations of rapid advancement, and its claims should not be attributed to every researcher who works in or near that environment. The text opens by saying that one can see the future first in San Francisco, describes a shift in what is discussed from billion-dollar to trillion-dollar clusters, and asserts that perhaps a few hundred people, most of them in San Francisco and the laboratories, have situational awareness of what is about to arrive [1]. It states that it is strikingly plausible that by 2027 models will be able to do the work of an AI researcher or engineer, and adds that this requires no belief in science fiction, only belief in straight lines on a graph. The author notes that he formerly worked at OpenAI and that the document rests on public information, his own ideas, general field knowledge, or what he calls San Francisco gossip. The essay takes no position on the forecasts, whose assessment would require evidence the three documents do not contain. It is interested in what the text shows about how a community forms its sense of what matters.

The relevant structure is epistemic. People who build and use a technology daily have access to information that outsiders lack, and the author’s report that the technology is useful in his daily tasks is an example of such access. The same conditions could make certain developments appear especially significant, if the community is selected for belief in the trajectory because those who judged it implausible were less likely to join, if its members converse mainly with each other, and if the institutions around them reward attention to the same questions. The forecast alone does not establish any of these conditions, and they are offered as hypotheses to be tested rather than as findings. The text itself describes its community as having once been derided as crazy and then vindicated by trusting trendlines, a narrative which, if accepted, would make continued trust in the trendlines feel like the lesson of experience. None of this refutes the forecasts. A community may be right because it is selected for early attention to what proves true. The point is that being closest to a technology and being most invested in a view of it are not separable features of such an environment, and that the formation of a research agenda within it deserves the same scrutiny that the agenda directs at others.

The reflexive turn follows from the experiment’s own concerns. Herdagdelen and colleagues argue that apparent consensus among experts who use models with similar priors may no longer

count as independent evidence, since agreement that arises from a shared source is weaker than agreement reached separately [9]. The argument applies to the community in which that very research is produced. Researchers who investigate whether people absorb an agent’s framing also make choices about which questions to ask and how to interpret answers, and those choices emerge inside an institution and a professional culture whose members may share assumptions that the research does not itself examine. This is no refutation of the findings, and it should be stated with the same care that is demanded elsewhere, since reflexive arguments can be turned on any field, including the commentator’s. It makes the formation of the agenda an appropriate subject of scrutiny rather than a given.

A concrete link between the documents makes the point sharper. The experiment’s discussion draws out an implication for automated safety and alignment research, noting that researchers have begun to use agents to run audits of frontier models, that a human sign-off assumes independent judgment, and that a misaligned agent could steer the research meant to detect its misalignment while sign-off makes it appear that people remain in charge [9]. The significance of the experiment does not depend on the forecast being realized. Existing uses of agents in consequential analytical work, which the authors cite, are enough to make influence on expert judgment worth investigating even if automated research never reaches the scale that *Situational Awareness* anticipates when its table of contents says that hundreds of millions of AGIs could automate AI research and compress a decade of algorithmic progress into a year [1]. Nor does an expectation that motivates an experiment make its findings circular, since evidence generated by a hypothesis can legitimately support that hypothesis, provided the measured outcome discriminates among alternatives. The sounder criticism is that the experiment tests a narrower proposition than the surrounding narrative. It shows that one-sided advice moves choices on underdetermined decisions in one task, whereas the picture of expert judgment across science and policy settled by the priors of a few models is a larger proposition, whose importance depends on the prevalence, consequences, and correctability of comparable influence in actual work. This is the process that the title names and the introduction describes.

8 Psychology, Measurement, and the Prestige of Neuroscience

The vocabulary of neuroscience carries an authority that behavioral description lacks, and its use in discussions of human-machine interaction illustrates a specific inferential problem more than it invites a general dismissal of psychology. Precise measurements can coexist with ambiguous constructs. An instrument may record electrical activity at a scalp with fine temporal resolution while the thing the researcher wishes to know, whether a person is thinking well, remains undefined by that measurement. Physiological language can make an interpretation appear more settled than the underlying inference warrants, because the measurement is exact even when the mapping from the measurement to the construct is not.

The position paper’s treatment of cognitive cost shows the structure. It states that the cognitive

cost is measurable and cites a study, reported in a preprint, which found that participants who used a language model for essay writing had significantly decreased brain connectivity [11, 10]. The inference chain implicit in this use runs from reduced connectivity during an assisted task, to cognitive debt, to degradation of the capacities that oversight needs. Each step requires its own support. Regional activity, functional connectivity, task engagement, subjective effort, and enduring ability are different things, and a difference in the first does not imply a difference in the last. Poldrack’s analysis of reverse inference in neuroimaging explains the difficulty: inferring a cognitive process from the activation of a region is only as strong as the selectivity of that region’s response to the process, and most regions respond to many processes [12]. A decrease in a neural measure does not carry an intrinsic negative meaning either. Practiced performance is commonly accompanied by reduced activation, and a task made easier by a tool may produce reduced activity because less work is being done, because the work is being done more efficiently, or because the person is attending to something else, and the neural record alone does not say which. The inference from a neural difference during assisted work to cognitive deterioration therefore needs independent behavioral and longitudinal support.

Such support exists in places, and the position paper cites some of it. The multicentre observational study by Budzyń and colleagues examined the unaided performance of clinicians in a specialist medical procedure after a period of exposure to AI assistance, and it addresses skill in a specific procedure by behavioral outcome rather than by neural proxy [3]. The experiment’s reference to work by Shen and Tamkin, in which software developers who used an assistant to learn a new library ended with weaker conceptual understanding, likewise bears on what persists after assistance, in a learning task where the question of retention is central [9]. These sources are stronger as evidence of deskilling than a connectivity measure, and they carry limits of their own: the first is observational, and the second concerns learning rather than expert performance. The position paper’s language of cognitive degradation, skill atrophy, and the destruction of oversight capability moves faster than this body of evidence supports, and the paper is candid that part of its base consists of self-reports and of findings it describes as emerging. The appropriate conclusion is that the evidence establishes plausible risk in specific settings and does not yet establish a general degradation.

The distinction also applies outside neuroscience, and the example that follows does a different job from the flow illustration in Section 9: that illustration shows that lower activity can accompany high engagement, whereas this one concerns an exact proxy that ranks representations differently from the capacity it is sometimes taken to indicate. Zou and colleagues compare compressed visual representations by feature-reconstruction error and by downstream multimodal understanding [15]. One residual-quantization variant produced about nine times the reconstruction error of ordinary vector quantization, 0.246 against 0.027, while achieving a higher average understanding score over eight benchmarks, 60.7 against 51.2, and a second variant whose reconstruction error was close to that of ordinary quantization scored substantially higher as well. The result concerns visual tokenizers and not human judgment, and it does not show that a discrete representation is semantically lossless, since an average over selected

benchmarks does not establish preservation of rare distinctions or support for untested decisions. It does establish the narrower point needed here: a precise proxy can rank representations in the opposite order from a downstream capacity. Reconstruction error was not designed as a measure of understanding, and the point is only that it answers a different question, as neural activity, experienced effort, and retained ability do. Precision about one does not establish its adequacy as a measure of another.

9 Flow, Automaticity, and the Possibility of Beneficial Offloading

The assumption that reduced effort signals reduced engagement has a plain counterexample in the division of attention within a single task. Removing minor decisions from a person’s workload may permit sustained attention to the larger undertaking, so that assistance diminishes engagement with some parts of a task in order to increase it with others. A programmer who no longer attends to the syntax of a library call may attend more closely to the design of the system, and an analyst relieved of boilerplate may spend the freed time on the identification strategy. Assistance might diminish engagement in some circumstances and enable it in others, and neither possibility can be inferred from reduced effort, fewer interventions, or a physiological measurement. The position paper’s own response to the objection that people adapt, which distinguishes searching for a document from authorizing a multistep plan, locates the difference in the structure of the task, and the structure of the task is likewise what determines when offloading is benign [11].

Offloading differs sharply between experts and learners, and the concern is strongest for learners. A novice who uses an agent to avoid the work through which skill is acquired may never acquire it, and the experiment’s reference to developers who learned a library with an assistant and ended with weaker conceptual understanding is evidence of this kind [9]. An expert who offloads previously mastered work may stand differently, provided the relevant skill remains recoverable and the assistance releases attention for other consequential parts of the task. The experiment’s participants were practitioners new to the domain, which places them between the two cases and limits what the study says about either.

The decisive question in both cases is what persists after assistance is withdrawn. Studies that remove the tool after a period of use and measure unaided performance on tasks of the same kind, and on tasks that require transfer, would address it directly. The experiment records, in the hand-coded sessions, that on 39 percent of decisions the participant neither asked about the decision nor took a side before the final choice [9]. That figure is compatible with a participant who judged the decision minor, one who attended to other parts of a sixty-minute task, and one who did not notice that a decision existed, and the study cannot say which, since it did not observe attention and the scenario carried no real policy stakes. Both papers point toward the measurement that would settle such cases, the position paper through its proposals for

domain skill maintenance exercises and the experiment through its call to estimate how an expert would have carried out a task without assistance.

The literature on flow offers a brief illustration and carries none of the weight of this argument. Csikszentmihalyi describes absorbed engagement in which self-monitoring recedes and performance is often at its best [4], Dietrich proposed that such states involve transient hypofrontality, a temporary down-regulation of explicit prefrontal processes while practiced processes continue [5, 6], and imaging work on experimentally induced flow reports differences in regional activity during absorbed engagement [13]. The hypothesis is contested and is not offered as an established counterexample. It illustrates only that lower activity in a region can accompany high engagement, which is the point made in Section 8 about connectivity, and the argument of this section would stand without it. The distinction between regional activity and connectivity applies here as it does in Section 8.

10 Oversight as an Organizational Problem

The position paper’s strongest contribution deserves its own treatment, and it does not depend on the broader claims of cognitive degradation. Assigning someone responsibility to supervise a fast, opaque system does not supply the time, information, authority, or expertise that supervision requires [11]. The paper documents how the volume of content an overseer must hold in mind, the multiple roles the user must play at once, and the poorly designed affordances of current systems make the nominal responsibility hard to discharge, and it observes that the onus of oversight falls almost entirely on users while mechanisms for exercising it are largely absent. This is a claim about the match between a responsibility and the resources that make it possible to meet it, and it remains important even if the more sweeping claims prove overstated. Regulatory texts that require effective human oversight presuppose that the human has the means to oversee, and the observation that the means are often not provided is a finding about organizations rather than about brains.

The same three-part test applies to the organizational argument, which fares better than the degradation claims on the first part and less well on the last. As to effect, the paper documents the mismatch between responsibility and resources through practitioner reports, workshop papers, an observational study of developers, and design analyses [11]. This is qualitative and observational evidence, consistent across sources, and suited to documenting a recurring mismatch across the settings studied, without establishing how common it is in a population or how often oversight fails. As to mechanism, the paper attributes failures to volume, role overload, and poorly designed affordances. These are plausible and in places directly documented, though the paper does not separate their contributions, so that a deployer cannot tell which remedy addresses the dominant cause. As to significance, the proposals presuppose that failures of oversight are consequential in deployed settings, and the evidence for their rate and cost is thin, which the paper acknowledges when it calls for system-level audits of whether existing agents

support the oversight they presume. The organizational claim is best read as a well-supported hypothesis about where to look, held with less evidence than its tone of urgency suggests.

The proposed interventions can be examined as proposals with costs. The paper offers an inventory that includes strategic friction, bounded autonomy and batch review, behavioral monitoring, canary tasks, trainings, enforced breaks and rotations, and role design, and it is careful to say that the inventory is not complete or definitive [11]. Additional friction and compulsory engagement have costs of their own, and the paper acknowledges that users may disprefer systems that reduce overreliance. The value of such measures should therefore be demonstrated rather than presumed, and the demonstration should use the right outcome. The relevant measure is whether an intervention improves detection and correction of consequential errors, which canary tasks and injected faults can test directly, and not whether the human appears busy. Effective oversight requires an appropriate arrangement of work, not a visibly engaged person.

The question of what to measure and how to arrange review has also been examined in discussions of the economics of research systems, which supply two considerations that bear on the position paper’s proposals [8]. The first concerns the unit of accounting. Evaluation by cost per task alone can favor inexpensive outputs without accounting for whether they survive verification, whereas cost per retained verified claim, which counts the effort of orchestration and checking, rewards outputs that end up linked to a source and independently confirmed. The distinction parallels the one drawn above between a busy reviewer and an effective one, since a count of tasks completed or reviews performed measures activity, while the quantity oversight is meant to produce is verified outcomes. The second concerns the form of correction. In work on model-to-model correction, constrained editing of a draft may reduce checking costs when the unchanged material has already been validated and the consequences of the edits remain local. The same structure applies by extension to a person reviewing an agent’s draft: restricting the editor to local changes makes each modification easier to inspect, because the reviewer compares a bounded difference, and it also anchors the result to the draft’s initial framing, which makes that framing harder to challenge. These consequences are compatible, and which dominates depends on what the review is meant to accomplish. Where review is meant to check execution within an accepted framing, constrained editing is an advantage. Where it is meant to question the framing itself, the same constraint may be a disadvantage, and independent regeneration or review from the sources may be required. A design for oversight of agent drafts should therefore state which of the two it intends.

The monitoring proposals expose a symmetry. The paper warns that approval signals treated as evidence of success can drift from the intended target of well-scrutinized correct action, so that optimizing for them rewards the conditions of ineffective oversight. Its own behavioral monitoring signatures, such as review time, override rate, and evidence-seeking, are proxies of the same kind, and once adopted as measures they invite the same drift, since reviewers measured on review duration may be encouraged to lengthen it. The remedy the paper gives for approval signals, auditing them against independent measures of oversight quality, applies

equally to its remedies. The motivating framework, dual-process theory with its partition between fast heuristic and slow deliberate thought, is a convenient heuristic whose status is disputed, and the recommendations do not need it, because the observations about time, information, and authority stand on their own.

The experiment’s conclusion converges on a related recommendation, which is that institutions decide deliberately which decisions need a person’s own informed judgment, give people the time and resources to form one, and surface key decision points as the work proceeds [9]. The two papers, despite their differences of genre and evidence, thus agree on the point that matters most for practice: the allocation of authority between people and machines should be something an institution chooses, with the resources to make the choice effective, rather than something it drifts into.

11 Conclusion: Evidence of Influence and Claims of Importance

The studio from the opening has produced an absorbing film, and the audience stayed. Designed experiences can produce effects worth studying, and corporate researchers can produce valuable findings. The instrument developed here separates what a study shows, why it shows it, and why it matters, and it can be run on a claim from each of the two documents that report evidence, while the forecast requires separate treatment.

The flip rate from the experiment is established as an effect: one-sided advice from a fluent agent moved choices on underdetermined decisions, and randomization supplies the causal reading. Its mechanism is open, since the design does not separate persuasion, reasonable delegation, convenience, and unexamined acceptance, and it contains no balanced-advice or human-adviser arm. Its significance depends on the prevalence, consequences, and correctability of comparable influence in actual work, none of which is established by the experimental effect alone.

The deskilling evidence in the position paper divides differently. As effect, its strongest items are behavioral, such as the multicentre clinical study by Budzyń and colleagues, which reports lower unaided detection rates after AI exposure under observational limits, while the connectivity finding, considered alone, does not establish a loss of ability. As mechanism, the paper’s account runs through offloading, automation bias, and fatigue, which are plausible and have not been separated empirically from a reallocation of attention to other parts of the task. As significance, the claim that oversight degrades the overseer in general moves ahead of evidence that supports plausible risk in specific settings.

The forecast in *Situational Awareness* has a different role and cannot be assessed through the same experimental categories. Its forecasts depend on the persistence of reported trends and on the relationship between those trends and future capabilities, questions this essay does not adjudicate. It serves here as a documented expression of expectations about the redistribution of intellectual work. Those expectations help explain why certain research questions appear

consequential, but the experimental findings do not independently establish the larger future the forecast describes.

Run in parallel, the test gives a consistent pattern for the two documents whose evidence this essay assesses: the effect is the best-supported element, the mechanism is the least examined, and the significance is asserted at a scale that the effect and mechanism do not yet carry. The gap between a precise measure and the capacity of interest is not peculiar to these documents, as the divergence between reconstruction error and downstream understanding reported by Zou and colleagues illustrates [15], and the instrument is accordingly not specific to neuroscience or to corporate research. The corporate research examined here is the site of the analysis and not its limit, since the findings are instances of a more general structure in which a precise measurement is read as evidence of a capacity or importance that it does not itself establish. Accepting assistance, exerting less effort, and contributing fewer original methodological choices do not individually settle whether a person is thinking well, and a company’s demonstration of influence does not establish that its technology is superior, that users have lost agency, or that its preferred account of the future is becoming inevitable. What the documents together specify is the evidence that is missing: studies that include balanced and human-adviser arms, that measure unaided performance after assistance, that test oversight by its success at finding consequential errors, that are reported as programs with their pilots and nulls, and that are open to replication by those who did not build the system. An industry’s expectations about its technology will continue to shape its research questions, as any field’s do, and the remedy lies in making the shaping visible and the evidence for each link in the chain separately assessable.

References

- [1] Aschenbrenner, L. *Situational Awareness: The Decade Ahead*. situational-awareness.ai, June 2024.
- [2] Bainbridge, L. Ironies of Automation. *Automatica*, 19(6):775–779, 1983.
- [3] Budzyń, K., et al. Endoscopist Deskilling Risk After Exposure to Artificial Intelligence in Colonoscopy: A Multicentre, Observational Study. *The Lancet Gastroenterology & Hepatology*, 10(10):896–903, 2025. Correction: *The Lancet Gastroenterology & Hepatology*, 10(11):e12, 2025. doi:10.1016/S2468-1253(25)00294-8.
- [4] Csikszentmihalyi, M. *Flow: The Psychology of Optimal Experience*. Harper & Row, New York, 1990.
- [5] Dietrich, A. Functional Neuroanatomy of Altered States of Consciousness: The Transient Hypofrontality Hypothesis. *Consciousness and Cognition*, 12(2):231–256, 2003.
- [6] Dietrich, A. Neurocognitive Mechanisms Underlying the Experience of Flow. *Consciousness and Cognition*, 13(4):746–761, 2004.

- [7] Endsley, M. R., and Kiris, E. O. The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors*, 37(2):381–394, 1995.
- [8] Flyxion. Research Beyond the Token Budget: Formal Reasoning, Representation, and the Economics of Computational Scale. 2026.
- [9] Herdağdelen, A., Joren, H., and Tamkin, A. Who’s Aligned to Whom? How AI Agents Shape Experts’ Research Judgment Calls. Manuscript, 2026. Preanalysis plan registered at osf.io/t3q59.
- [10] Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. Your Brain on ChatGPT: Accumulation of Cognitive Debt When Using an AI Assistant for Essay Writing Task. arXiv preprint arXiv:2506.08872, 2025.
- [11] Mitchell, M., Ghosh, A., and Passi, S. AI Agents Push Humans Out of the Loop. arXiv preprint arXiv:2608.23642v1, August 2026.
- [12] Poldrack, R. A. Can Cognitive Processes Be Inferred from Neuroimaging Data? *Trends in Cognitive Sciences*, 10(2):59–63, 2006.
- [13] Ulrich, M., Keller, J., Hoenig, K., Waller, C., and Grön, G. Neural Correlates of Experimentally Induced Flow Experiences. *NeuroImage*, 86:194–202, 2014.
- [14] Rosenthal, R. The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin*, 86(3):638–641, 1979.
- [15] Zou, J., Li, X., Yue, Z., Huang, X., Luo, Y., Yan, W., Li, H., Liu, W., and Wang, X. Can Discrete Tokens Serve as Semantically Lossless Representations of Visual Signals? arXiv preprint arXiv:2610.37662, 2026.